MASTER EN SCIENCES MATHÉMATIQUES

Méthodes de points intérieurs en programmation du cône du second ordre: une approche par algèbre de Jordan

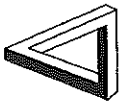
Russo, Frank

Award date:
2003

FUNDP
Faculté des Sciences
Département de Mathématique

Rempart de la Vierge, 8
B-5000 Namur Belgique

Méthodes de points intérieurs en programmation du cône du second ordre. Une approche par algèbre de Jordan



Mémoire présenté pour l'obtention
du grade de
Licencié en Sciences Mathématiques
par

Frank Russo

Promoteur : J-J. Strodiot

Année Académique 2002-2003

Je tiens à remercier principalement mon promoteur, Monsieur J-J. Strodiot, pour son aide, sa disponibilité et l'attention qu'il a portée à ce mémoire.

Je remercie également les professeurs F. Alizadeh, F. Callier et J. Sturm pour leur aide et leurs conseils.

Enfin, je remercie toutes les personnes qui ont contribué à ma formation et toutes celles qui m'ont soutenu, de près ou de loin, durant ces quatres années d'études.

Résumé

L'objectif recherché dans ce mémoire est d'étudier la programmation du cône du second ordre et d'aboutir au développement de méthodes de points intérieurs pour la résolution de problèmes du cône du second ordre (problèmes SOCP). Un problème SOCP consiste à minimiser une fonction linéaire sur l'intersection d'un ensemble affine et d'un produit cartésien de cônes du second ordre. En premier lieu, nous donnons les définitions de base liées au cône du second ordre, pour ensuite présenter quelques exemples de problèmes de natures différentes pouvant se modéliser en problème SOCP. Après cela, nous introduisons une algèbre "taillée sur mesure" pour les problèmes SOCP en étudiant le concept d'algèbre de Jordan. La compréhension de cette algèbre est essentielle pour aborder, par la suite, les notions de dualité faible et forte, conditions de complémentarité, non-dégénérescence et complémentarité stricte. Nous terminerons par le développement de méthodes de points intérieurs utilisant des algorithmes de suivi de chemins de complexité polynomiale. Enfin, nous donnons en annexe la preuve de résultats techniques ainsi que la description complète de quelques exemples concrets de problèmes SOCP accompagnée de résultats numériques.

Abstract

The aim of this work is to survey second order cone programming (SOCP) and to end up with the development of interior point methods for the solution of second order cone problems. A SOCP problem is a convex optimization problem in which a linear function is minimized over the intersection of an affine set and the cartesian product of second order cones. First, we give the basic definitions connected with the second order cone and present several examples of problems arising from various areas and which can be cast as SOCP problems. Next, we introduce a "tailor-made" algebra for SOCP by studying the concept of Jordan algebra . The comprehension of this algebra is essential to move on to the notions of weak and strong duality, complementarity conditions, non-degeneracy and strict complementarity. We end up with the development of primal-dual interior point methods using polynomial complexity path following algorithms. In the appendices, we give the proofs of some technical results and the complete description of some concrete engineering SOCP problems, followed by numerical results.

Table des matières

Introduction 3

1 Le cône du second ordre 6
1.1 Notations et définitions 6
1.2 Caractérisation du cône du second ordre 12

2 Problèmes SOCP 14
2.1 Formulation de problèmes SOCP 14
2.1.1 Programmation linéaire 14
2.1.2 Programmation quadratique 14
2.1.3 Problèmes de minimisation de normes 16
2.1.4 Problèmes avec contraintes hyperboliques 17
2.2 Problèmes robustes 24
2.2.1 Programmation linéaire robuste 24
2.2.2 Moindres carrés robustes 26

3 Algèbre du cône du second ordre 28
3.1 Algèbre de Jordan 28
3.1.1 Notion d'algèbre 28
3.1.2 Algèbres de Jordan 32
3.2 Algèbre SOCP 33
3.2.1 Un cadre particulier 33
3.2.2 Décomposition spectrale 35
3.2.3 Notions de degré et rang. 39
3.2.4 La représentation quadratique 42
3.2.5 Commutativité et structure de Jordan 45
3.3 Propriétés de Q_x 48

4 Dualité pour SOCP 57
4.1 Dualité et complémentarité 57
4.1.1 Dualité faible, semi-forte et forte 57
4.1.2 Complémentarité 60
4.2 Non-dégénérescence et complémentarité stricte 63
4.2.1 Non-dégénérescence 63

4.2.2 Complémentarité stricte 69
4.3 Non-singularité du jacobien 73

5 Méthodes de points intérieurs **79**
5.1 Introduction 79
5.2 Barrière logarithmique 80
5.3 Trajectoire centrale et direction de Newton 82
5.4 Changements d'échelle 84
5.5 Directions commutatives 86
5.6 Algorithmes de suivi de chemins 88
5.6.1 Lemmes techniques 89
5.6.2 Bornes sur δ_x et δ_z 96
5.6.3 Bornes sur le nombre de conditionnement de G 99
5.6.4 Complexité des algorithmes 101

Conclusion **104**

Annexes **107**

A Preuve de la $\mathcal{S}$-procédure **107**
A.1 Lemmes préparatoires 107
A.2 Preuve de la $\mathcal{S}$-procédure 110

B Applications **113**
B.0.1 Synthèse d'un tableau d'antennes 113
B.0.2 Conception d'une topologie en trellis 120
B.0.3 Optimisation d'un portefeuille 124
B.0.4 Equilibre d'un système de ressorts linéaires par morceaux. 126
B.1 Tests numériques avec SeDuMi. 129

Bibliographie **131**

Introduction

Ce mémoire est consacré à l'étude de la programmation du cône du second ordre . Les problèmes de programmation du cône du second ordre (en abrégé, problèmes SOCP - de l'anglais *Second Order Cone Programming*), sont des problèmes d'optimisation convexe dans lesquels une fonction linéaire est minimisée sur l'intersection d'un sous-espace affine et d'un produit cartésien de cônes du second ordre. Dans la littérature, ce cône est aussi appelé cône quadratique, cône de *Lorentz* ou encore *ice-cream cone*. Ces problèmes ont la forme standard suivante :

$$\begin{cases} \min & c^T x \\ \text{s.c.} & Ax = b \\ & x \in \mathcal{Q} \end{cases}$$

où $\mathcal{Q}$ désigne le cône du second ordre (voir plus tard pour les définitions) ou un produit cartésien de ce cône. La contrainte $x \in \mathcal{Q}$ est souvent notée $x \succcurlyeq_{\mathcal{Q}} 0$. Comme nous le verrons, les programmes linéaires, les programmes quadratiques convexes à contraintes linéaires ainsi que les programmes quadratiques à contraintes quadratiques convexes pourront tous être formulés comme des problèmes SOCP ; ce sera le cas également pour d'autres types de problèmes. La raison principale qui incite à l'étude des problèmes SOCP est qu'ils permettent de modéliser des applications provenant d'une large variété de domaines tels que, par exemple, l'ingénierie, le contrôle, la finance, l'optimisation robuste et l'optimisation combinatoire.

Un des problèmes d'optimisation les plus célèbres est le problème appelé problème de *Fermat-Weber*. Il pose la question suivante : soit un ensemble de N points d_i situés dans un repère euclidien et représentant par exemple des points de vente. Où placer le point x (par exemple, un entrepôt) tel que la somme des distances de chacun des points d_i à x soit minimale ? Cela revient à résoudre le problème non-linéaire suivant :

$$\min_x \sum_{i=1}^{N} \|d_i - x\|, \quad \text{où } d_i,\ i = 1, \ldots, N \text{ est fixé.}$$

Comme il n'est pas possible de mettre ce problème sous la forme d'un programme linéaire, il faut penser à une autre stratégie. Une façon de procéder

est de modéliser ce problème comme un problème SOCP standard. L'idée est d'utiliser une fonction objectif correspondant à une somme de fonctions linéaires qu'il faut minimiser, et d'imposer à chaque terme $\|d_i - x\|$ de rester inférieur à chacune de ces fonctions linéaires.

De plus, les problèmes SOCP sont considérés comme des cas particulier de problèmes de programmation semi-définie; ces derniers correspondent à des problèmes d'optimisation sur l'intersection d'un sous-espace affine et du cône des matrices semi-définies positives. Au vu des observations faites jusqu'ici, la programmation du cône du second ordre (que noterons désormais SOCP, en abrégé) se situe entre la programmation linéaire, quadratique et semi-définie. Tout comme des problèmes de programmtion linéaire, quadratique et semi-définie, les problèmes SOCP peuvent être résolus par des algorithmes de points intérieurs de complexité polynomiale. D'un point de vue algorithmique, l'effort par itération requis par ces méthodes de points intérieurs est plus important que pour des problèmes linéaires ou quadratiques mais moindre que pour des problèmes de programmation semi-définie de même taille. Remarquons également que la région admissible pour des problèmes SOCP n'est pas polyhédrale, c'est pourquoi l'idée de développer des méthodes similaires à celle du simplexe, pour ces problèmes, serait une idée peu recommandable.

Nous allons voir qu'il existe une théorie qui est basée sur les algèbres de Jordan euclidiennes qui permet de faire le lien entre la programmation linéaire, semi-définie et SOCP. Le livre de Faraut et Koranyi [8] couvre l'étude des cônes symétriques et montre comment de tels cônes peuvent être munis d'une algèbre de Jordan. Nous nous pencherons plus précisément sur l'étude de l'algèbre de Jordan propre au cône du second ordre qui sera un passage nécessaire pour pouvoir aborder, par la suite, les méthodes de points intérieurs.

De nos jours, la programmation du cône du second ordre fait l'objet de nombreuses recherches dans tous les domaines où elle est utilisée. En particulier, beaucoup de mathématiciens s'y intéressent; c'est le cas notamment de Nesterov et Nemirovski qui ont montré que des méthodes barrière s'appliquent aux problèmes SOCP, conduisant à des algorithmes de points intérieurs de complexité $\sqrt{r}$ pour des problèmes comportant r inégalités du cône du second ordre. Nemirovksi et Scheinberg ont montré que les méthodes primales ou duales de points intérieurs développées en programmtion linéaire s'étendent facilement à SOCP. Quant aux méthodes primales-duales de points intérieurs pour les problèmes SOCP, elles ont été introduites pour la première fois par Nesterov et Todd qui sont à l'origine de la méthode NT. Ces méthodes primales-duales se sont avérées plus robustes d'un point de vue numérique que des méthodes uniquement primales ou duales. D'autres encore comme Adler, Alizadeh et Schmieta ont étudié la relation entre les problèmes SOCP et semi-définis et ont adapté la méthode $XZ + ZX$ aux problèmes SOCP. Ils ont ensuite développé des conditions de non-dégénérescence pour SOCP ainsi qu'une implémentation numériquement stable de la méthode $XZ + ZX$. Par la suite, Monteiro et Tsuchiya ont prouvé

la compléxité polynomiale de cette méthode.

Quelques packages permettant à des softwares de résoudre (entre autres) des problèmes SOCP sont actuellement disponibles gratuitement sur Internet. C'est le cas, par exemple, du package SDPpack, qui implémente la méthode $XZ+ZX$ mentionnée juste-avant, ainsi que du package SeDuMi [1] développé par Jos Sturm et qui est basé sur la méthode NT.

Dans le chapitre 1 nous donnons les définitions de base et fixons les formes standard pour le primal et le dual. Après cela, nous donnerons quelques exemples de problèmes SOCP et nous verrons comment des problèmes de natures différentes peuvent être mis sous la forme d'un problème SOCP ; c'est ce dont nous allons discuter dans le chapitre 2. Le chapitre 3 sera consacré à l'étude de l'algèbre du cône du second ordre qui est un cas particulier d'algèbre de Jordan. Cette étude est essentielle pour l'analyse des algorithmes de points intérieurs pour SOCP. Les notions de dualité, de conditions de complémentarité et de non-dégénérescence seront abordées dans le chapitre 4. Enfin, le chapitre 5 aura pour objectif d'établir les algorithmes de points intérieurs correspondant, en réalité, à des algorithmes de suivi de chemins.

[1] SDPPack ainsi que ses guides d'installation et d'utilisation sont disponibles à l'adresse http ://www.cs.nyu.edu/cs/faculty/overton/sdppack. Pour le package SeDuMi , l'adresse est : http ://fewcal.kub.nl/sturm/software/sedumi.html.

Chapitre 1

Le cône du second ordre

1.1 Notations et définitions

Les problèmes d'optimisation que nous allons traiter concerneront, de façon générale, des vecteurs partitionnés en blocs où chaque bloc est un vecteur indexé à partir de 0. Nous utiliserons des lettres en gras $\mathbf{x}, \mathbf{c}$ etc. pour de tels vecteurs et $\mathbf{x}_i$ représentera le i-ème bloc de $\mathbf{x}$. La notation x_j ou x_{ij} sera utilisée pour faire référence à la j-ème coordonnée du vecteur ou du bloc correspondant. Les vecteurs $\mathbf{0}$ et $\mathbf{1}$ désigneront respectivement le vecteur constitué de 0 à chaque entrée et le vecteur constitué de 1 à chaque entrée. Les matrices seront notées de façon classique par des majuscules et nous ne specifierons pas leurs dimensions lorsque le contexte sera suffisamment clair.
Souvent, nous devrons concaténer des vecteurs ou des matrices. Pour mentionner ces concaténations, nous utiliserons la notation "," pour concaténer des vecteurs ou des matrices sur une ligne, ainsi que la notation " ;" pour les concaténer sur une colonne. Ainsi, pour les vecteurs $\mathbf{x}$, $\mathbf{y}$ et $\mathbf{z}$ par exemple, nous avons les trois notations équivalentes suivantes :

$$\begin{pmatrix} \mathbf{x} \\ \mathbf{y} \\ \mathbf{z} \end{pmatrix} = (\mathbf{x}^T, \mathbf{y}^T, \mathbf{z}^T)^T = (\mathbf{x}; \mathbf{y}; \mathbf{z}).$$

Si $\mathcal{A} \subseteq I\!R^k$ et $\mathcal{B} \subseteq I\!R^l$ alors

$\mathcal{A} \times \mathcal{B} = \{(\mathbf{x}; \mathbf{y}) \,|\, \mathbf{x} \in \mathcal{A} \text{ et } \mathbf{y} \in \mathcal{B}\}$ désigne leur produit cartésien.

Pour deux matrices A et B, nous définissons la somme directe de A et B par :

$$A \oplus B = \begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix}.$$

Dans ce qui va suivre, nous allons nous intéresser exclusivement à des cônes $\mathcal{K} \subseteq I\!R^n$ fermés, pointés, convexes et d'intérieur non vide, que l'on munira d'un ordre partiel.

Définition 1.1.1 *Un ensemble $\mathcal{K}$ est un cône si il est fermé sous multiplication par un scalaire positif ou nul, c'est-à-dire si :*

$$\mathbf{x} \in \mathcal{K} \text{ et } \alpha \geq 0 \Rightarrow \alpha\mathbf{x} \in \mathcal{K}.$$

Ce cône est dit pointé si $\mathcal{K} \cap (-\mathcal{K}) = \{0\}$ où $-\mathcal{K} = \{x \mid -x \in \mathcal{K}\}$.

Il est possible, comme pour l'espace $\mathbb{R}^n$, de définir un ordre partiel sur ces cônes qui sera représenté par la relation $\succcurlyeq_{\mathcal{K}}$ et défini de la façon suivante :

$$\mathbf{x} \succcurlyeq_{\mathcal{K}} \mathbf{y} \Leftrightarrow \mathbf{x} - \mathbf{y} \in \mathcal{K} \quad \text{et} \quad \mathbf{x} \succ_{\mathcal{K}} \mathbf{y} \Leftrightarrow \mathbf{x} - \mathbf{y} \in \operatorname{int}\mathcal{K}. \tag{1.1}$$

Grâce à l'hypothèse faite sur les cônes avec lesquels nous allons travailler, cette relation satisfait bien les propriétés nécessaires pour être un ordre partiel. En effet,

- *reflexivité :* nous avons bien $\mathbf{x} \succcurlyeq_{\mathcal{K}} \mathbf{x}\ \forall \mathbf{x} \in \mathcal{K}$ car $0 \in \mathcal{K}$.
- *antisymétrie :* si $\mathbf{x} \succcurlyeq_{\mathcal{K}} \mathbf{y}$ et $\mathbf{y} \succcurlyeq_{\mathcal{K}} \mathbf{x}$ alors $\mathbf{x} - \mathbf{y} \in \mathcal{K}$ et $-(\mathbf{x} - \mathbf{y}) \in \mathcal{K}$. Donc, comme $\mathcal{K}$ est pointé, $\mathbf{x} = \mathbf{y}$.
- *transitivité :* supposons que $\mathbf{x} \succcurlyeq_{\mathcal{K}} \mathbf{y}$ et $\mathbf{y} \succcurlyeq_{\mathcal{K}} \mathbf{z}$. Alors $\mathbf{x} \succcurlyeq_{\mathcal{K}} \mathbf{z}$ car $\mathbf{x} - \mathbf{z} = 2\{\frac{1}{2}(\mathbf{x} - \mathbf{y}) + \frac{1}{2}(\mathbf{y} - \mathbf{z})\} \in \mathcal{K}$ puisque $\mathcal{K}$ est un cône convexe.

De plus, cet ordre est compatible avec les opérateurs linéaires dans le sens où :

$$\begin{array}{lcl} \mathbf{a} \succcurlyeq_{\mathcal{K}} \mathbf{b} \text{ et } \lambda \geq 0 & \Rightarrow & \lambda\mathbf{a} \succcurlyeq_{\mathcal{K}} \lambda\mathbf{b} \\ \mathbf{a} \succcurlyeq_{\mathcal{K}} \mathbf{b} \text{ et } \mathbf{c} \succcurlyeq_{\mathcal{K}} \mathbf{d} & \Rightarrow & \mathbf{a} + \mathbf{c} \succcurlyeq_{\mathcal{K}} \mathbf{b} + \mathbf{d}. \end{array}$$

Ces deux implications découlent directement du fait que $\mathcal{K}$ est un cône convexe pointé. Notons que les relations $\preccurlyeq_{\mathcal{K}}$ et $\prec_{\mathcal{K}}$ sont définies à partir de $\succcurlyeq_{\mathcal{K}}$ et $\succ_{\mathcal{K}}$:

$$\mathbf{x} \preccurlyeq_{\mathcal{K}} (\prec_{\mathcal{K}})\mathbf{y} \Leftrightarrow \mathbf{y} \succcurlyeq_{\mathcal{K}} (\succ_{\mathcal{K}})\mathbf{x}$$

Définition 1.1.2 *A chaque cône $\mathcal{K} \subseteq \mathbb{R}^n$ on associe son cône dual défini par*

$$\mathcal{K}^* = \{\mathbf{z} \in \mathbb{R}^n \mid \mathbf{x}^T\mathbf{z} \geq 0\ \forall \mathbf{x} \in \mathcal{K}\}.$$

En particulier, nous obtenons que pour $\{0\} \subseteq \mathbb{R}^n$, $\{0\}^* = \mathbb{R}^n$ ainsi que $(\mathbb{R}^n)^* = \{0\}$. Dans un cadre tout-à-fait général, $\mathcal{K}$ est le produit cartésien de plusieurs cônes : $\mathcal{K} = \mathcal{K}_{n_1} \times \ldots \times \mathcal{K}_{n_r}$, où chaque $\mathcal{K}_{n_i} \subseteq \mathbb{R}^{n_i}$. Dans ce cas, les vecteurs $\mathbf{x}$, $\mathbf{c}$ et $\mathbf{z}$ ainsi que la matrice A sont partitionnés en accord avec le produit cartésien, i.e. :

$$\begin{array}{ll} \mathbf{x} = (\mathbf{x}_1; \ldots; \mathbf{x}_r) & \text{où } \mathbf{x}_i \in I\!R^{n_i}, \\ \mathbf{z} = (\mathbf{z}_1; \ldots; \mathbf{z}_r) & \text{où } \mathbf{z}_i \in I\!R^{n_i}, \\ \mathbf{c} = (\mathbf{c}_1; \ldots; \mathbf{c}_r) & \text{où } \mathbf{c}_i \in I\!R^{n_i} \\ A = (A_1, \ldots, A_r) & \text{où chaque } A_i \in I\!R^{m \times n_i}. \end{array} \tag{1.2}$$

Par la suite, r désignera le nombre de blocs du vecteur $\mathbf{x}$, $n = \sum_{i=1}^{r} n_i$ la dimension du problème et m le nombre de lignes de chaque A_i. Pour chaque vecteur $\mathbf{x} \in I\!R^n$ constitué d'un seul bloc et indexé à partir de 0, nous noterons par $\bar{\mathbf{x}} \in I\!R^{n-1}$ le sous-vecteur constitué des composantes indexées de 1 à $n-1$; ainsi $\mathbf{x} = (x_0; \bar{\mathbf{x}})$. Nous noterons également $\hat{\mathbf{x}} = (0; \bar{\mathbf{x}})$. De façon similaire, pour une matrice $A \in I\!R^{m \times n}$ dont les colonnes sont indexées à partir de 0, $\bar{A}$ désigne la sous-matrice de A constituée des colonnes 1 jusque $n-1$.
Ces notations étant fixées nous donnons à présent la définition du cône du second ordre. Nous nous plaçons dorénavant dans le cadre où chaque cône $\mathcal{K}_{n_i}$ est le cône du second ordre.

Définition 1.1.3 *Le cône du second ordre (ou cône quadratique, ou cône de Lorentz, ou encore ice-cream cone) de dimension n est défini par :*

$$\mathcal{Q}_n = \{\mathbf{x} = (x_0; \bar{\mathbf{x}}) \in I\!R^n \,|\, x_0 \geq \|\bar{\mathbf{x}}\|\}.$$

Le qualificatif "du second ordre" sera justifiée au chapitre 3 et l'appellation alternative "cône de *Lorentz*" provient du fait que le physicien *Lorentz* a utilisé ce cône avec $n = 3$ en théorie de la relativité ; la quatrième dimension correspondait au temps. Dans la définition, $\|.\|$ correspond à la norme euclidienne. Si la dimension n du cône du second ordre est évidente par le contexte ou n'a pas d'importance, l'indice n dans $\mathcal{Q}_n$ sera négligé.

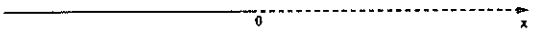


figure 1.1: A une dimension, le cône du second ordre (en pointillés) coïncide avec $I\!R_+$.

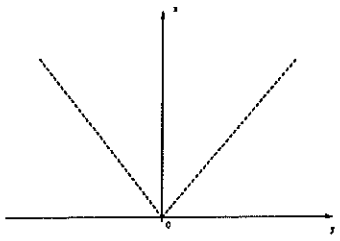


figure 1.2: Le cône du second ordre de dimension 2 dont les bords sont en pointillés.

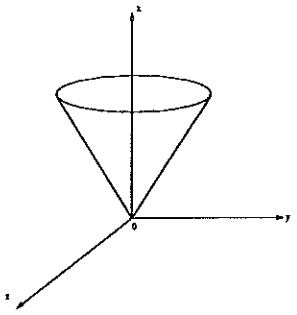


figure 1.3: Le cône du second ordre de dimension 3.

Signalons que pour la clarté et afin d'éviter une lourdeur dans les notations, nous allons souvent considérer le cas où chaque vecteur $\mathbf{x}$ comporte un seul bloc, c'est-à-dire que le produit cartésien de cônes du second ordre est constitué d'un seul cône. Lorsque le partitionnement en plusieurs blocs sera nécessaire, nous considèrerons $\mathcal{Q}$ comme un produit cartésien.

A présent, pour que la relation d'ordre définie en (1.1) soit bien définie dans ce cadre, il est important de s'assurer que $\mathcal{Q}$ est bien un cône fermé, pointé, convexe et d'intérieur non vide.
En effet, en remarquant que l'intérieur de $\mathcal{Q}$ est défini par

$$\operatorname{int}\mathcal{Q} = \{\mathbf{x} = (x_0; \bar{\mathbf{x}}) \in \mathcal{Q} \,|\, x_0 > \|\bar{\mathbf{x}}\|\},$$

il est immédiat que cet ensemble est non vide puisqu'il contient le vecteur $(1; \mathbf{0})$. Le caractère fermé de $\mathcal{Q}$ provient immédiatement de la continuité de la norme euclidienne. Supposons à présent que $\mathbf{x}$ et $-\mathbf{x} \in \mathcal{Q}$. Nous obtenons alors :

$$0 \leq \|\bar{\mathbf{x}}\| \leq x_0 \leq -\|\bar{\mathbf{x}}\| \leq 0 \quad \Rightarrow \quad \mathbf{x} = (x_0; \bar{\mathbf{x}}) = 0$$

d'où, $\mathcal{Q}$ est pointé. Enfin, avec $\mathbf{x}$ et $\mathbf{y} \in \mathcal{Q}$ et $t \in]0, 1[$ nous obtenons :

$$tx_0 + (1-t)y_0 \geq t\|\bar{\mathbf{x}}\| + (1-t)\|\bar{\mathbf{y}}\| \geq \|t\bar{\mathbf{x}} + (1-t)\bar{\mathbf{y}}\|$$

qui montre que $\mathcal{Q}$ est convexe.

Voici à présent un premier résultat intéressant concernant $\mathcal{Q}$:

Théorème 1.1.1 *Le cône du second ordre $\mathcal{Q}$ est auto-dual, i.e. $\mathcal{Q} = \mathcal{Q}^*$.*

Preuve :
1)$\underline{\mathcal{Q} \subseteq \mathcal{Q}^*}$

Soit $\mathbf{x} \in \mathcal{Q}$. En utilisant l'inégalité de Cauchy-Schwarz, nous obtenons :

$\forall\, \mathbf{z} \in \mathcal{Q}, \quad \mathbf{x}^T\mathbf{z} = x_0 z_0 + \bar{\mathbf{x}}^T\bar{\mathbf{z}} \geq \|\bar{\mathbf{x}}\|\|\bar{\mathbf{z}}\| - \|\bar{\mathbf{x}}\|\|\bar{\mathbf{z}}\| = 0$. D'où, $\mathbf{x} \in \mathcal{Q}^*$.

2)$\underline{\mathcal{Q}^* \subseteq \mathcal{Q}}$

Soit $\mathbf{x} \in \mathcal{Q}^*$. Nous allons considérer séparement les cas $\bar{\mathbf{x}} \neq 0$ et $\bar{\mathbf{x}} = 0$.
Supposons d'abord que $\bar{\mathbf{x}} \neq 0$. En posant $\mathbf{z} = \left(1; -\frac{\bar{\mathbf{x}}}{\|\bar{\mathbf{x}}\|}\right)$ on s'aperçoit que $\mathbf{z} \in \mathcal{Q}$.
De plus, $x_0 - \|\bar{\mathbf{x}}\| = x_0 z_0 + \bar{\mathbf{x}}^T\bar{\mathbf{z}} = \mathbf{x}^T\mathbf{z} \geq 0$. Nous obtenons bien $x_0 \geq \|\bar{\mathbf{x}}\|$, d'où $\mathbf{x} \in \mathcal{Q}$.
Supposons ensuite que $\bar{\mathbf{x}} = 0$. En posant $\mathbf{z} = (z_0; \bar{\mathbf{z}}) = (1; \mathbf{0})$ nous avons aussi $\mathbf{z} \in \mathcal{Q}$. Comme $\mathbf{x} \in \mathcal{Q}^*$ nous avons $\mathbf{x}^T\mathbf{z} \geq 0$. Si, par l'absurde, nous avions $x_0 < \|\bar{\mathbf{x}}\|$ alors nous aurions :

$0 \leq \mathbf{x}^T\mathbf{z} = x_0 z_0 + \bar{\mathbf{x}}^T\bar{\mathbf{z}} = x_0 < \|\bar{\mathbf{x}}\| = 0$, ce qui est impossible, d'où $\mathbf{x} \in \mathcal{Q}$.

□

Du fait du rôle particulier joué par la coordonnée x_0 dans le cône du second ordre, il est utile de définir la matrice de réflexion par rapport à l'axe des coordonnées x_0 :

$$R_n = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & -1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & -1 \end{pmatrix} \in I\!R^{n\times n},$$

ainsi que le vecteur

$$\mathbf{e}_n = (1; \mathbf{0}) \in I\!R^n.$$

Ici aussi nous négligerons l'indice n dans R_n et $\mathbf{e}_n$ lorsque celui sera évident d'après le contexte.
Nous donnons à présent les formes standard pour le primal et le dual d'un problème SOCP.

Déflnition 1.1.4 *Les formes standard primale et duale pour un problème de programmation du cône du second ordre sont :*

Primal		**Dual**	
min	$\mathbf{c}_1^T\mathbf{x}_1 + \ldots + \mathbf{c}_r^T\mathbf{x}_r$	max	$b^T y$
s.c.	$A_1\mathbf{x}_1 + \ldots + A_r\mathbf{x}_r = b$	*s.c.*	$A_i^T y + \mathbf{z}_i = \mathbf{c}_i \quad i = 1, \ldots, r$
	$\mathbf{x}_i \succcurlyeq_{\mathcal{Q}} \mathbf{0} \quad i = 1, \ldots, r$		$\mathbf{z}_i \succcurlyeq_{\mathcal{Q}} \mathbf{0} \quad i = 1, \ldots, r.$

(1.3)

Les inégalités qui figurent dans (1.3) seront appelées inégalités du cône du second ordre.

Remarquons que dans les inégalités du cône du second ordre du dual, l'indice $\mathcal{Q}$ remplace l'indice $\mathcal{Q}^*$ que l'on utilise en toute généralité pour un problème de programmation conique dont le cône sous-jacent n'est pas forcément auto-dual.

A ce stade, nous faisons les deux hypothèses suivantes au sujet de la paire primale-duale (1.3) :

Hypothèse 1. Les lignes de la matrice $A = (A_1, \ldots, A_r) \in I\!R^{m \times n}$ sont linéairement indépendantes.

Hypothèse 2. Les problèmes primal et dual sont tous deux strictement admissibles ; c'est-à-dire qu'il existe un vecteur $\mathbf{x} = (\mathbf{x}_1; \ldots; \mathbf{x}_r)$ admissible pour le primal tel que $\mathbf{x}_i \succ_{\mathcal{Q}} \mathbf{0}$ pour $i = 1, \ldots, r$, et il existe une paire $(y, \mathbf{z})$ admissible pour le dual telle que $\mathbf{z}_i \succ_{\mathcal{Q}} \mathbf{0}$, pour $i = 1, \ldots, r$. Une telle hypothèse est tout à fait naturelle lorsque l'on a pour objectif de développer des méthode de points intérieurs. Nous verrons plus tard les complications qui peuvent se produire lorsque cette hypothèse n'est pas vérifiée.

La proposition qui va suivre caractérise le problème dual associé au dual dans (1.3).

Proposition 1.1.1 *Le problème dual associé au dual dans (1.3) correspond au primal.*

Preuve : Tout d'abord, observons que le dual dans (1.3) peut être réécrit sous une forme analogue à celle du primal

$$\begin{cases} \min & -b^T y + \mathbf{0}^T \mathbf{z} \\ \text{s.c.} & [A^T \; I] \begin{pmatrix} y \\ \mathbf{z} \end{pmatrix} = \mathbf{c} \\ & (y, \mathbf{z}) \in I\!R^m \times \mathcal{Q}. \end{cases}$$

Puisque $\mathcal{Q}^* = \mathcal{Q}$ et $(I\!R^m)^* = \{0\}$, le cône dual à $I\!R^m \times \mathcal{Q}$ est $\{0\} \times \mathcal{Q}$ et le problème dual du dual dans (1.3) s'écrit

$$\begin{cases} \max & \mathbf{c}^T \mathbf{s} \\ \text{s.c.} & (-b; \mathbf{0}) - (A\mathbf{s}; \mathbf{s}) \in \{0\} \times \mathcal{Q}, \end{cases}$$

c'est-à-dire,

$$\begin{cases} \max & \mathbf{c}^T \mathbf{s} \\ \text{s.c.} & A\mathbf{s} + b = 0 \\ & -\mathbf{s} \in \mathcal{Q}. \end{cases}$$

En posant $\mathbf{x} := -\mathbf{s}$, nous retrouvons bien le primal de (1.3).

□

1.2 Caractérisation du cône du second ordre

Il existe, pour chaque vecteur $\mathbf{x} \in I\!R^n$, une matrice dans $I\!R^{n\times n}$ qui jouera un rôle essentiel dans tout ce qui va suivre. Elle est notée $Arw(\mathbf{x})$ (de l'anglais *arrow=flèche*) et est définie par :

$$Arw(\mathbf{x}) = \begin{pmatrix} x_0 & \bar{\mathbf{x}}^T \\ \bar{\mathbf{x}} & x_0 I \end{pmatrix}.$$

Sa structure en flèche vient du fait que la première colonne est constitué du vecteur $\mathbf{x}$, la première ligne correspond à $\mathbf{x}^T$, la digonale ne contient que la composante x_0 et partout ailleurs la matrice est constituée de 0.
Enonçons à présent un premier résultat intéressant concernant $Arw(\mathbf{x})$.

Proposition 1.2.1 $\forall\, \mathbf{x} \in I\!R^n$ *nous avons :*

$$Arw(\mathbf{x}) \succcurlyeq 0 (\succ 0) \Longleftrightarrow \mathbf{x} \succcurlyeq_{\mathcal{Q}} 0 (\succ_{\mathcal{Q}} 0).$$

Preuve : Les valeurs propres de $Arw(\mathbf{x})$ peuvent être calculées facilement en fonction des valeurs $\mathbf{x}$ grâce à sa structure particulière. Pour cela, nous allons montrer par récurrence sur la dimension de $Arw(\mathbf{x})$, que le déterminant de $Arw(\mathbf{x}) - \lambda I$ est égal à :

$$p_n(\mathbf{x}, \lambda) = (x_0 - \lambda)^{n-2}(\lambda^2 - 2x_0\lambda + x_0^2 - \|\bar{\mathbf{x}}\|^2) \tag{1.4}$$

ou encore

$$p_n(\mathbf{x}, \lambda) = (x_0 - \lambda)^{n-2}(\lambda - (x_0 + \|\bar{\mathbf{x}}\|))(\lambda - (x_0 - \|\bar{\mathbf{x}}\|))$$

où n est la dimension de $\mathbf{x}$.
Pour $n = 1$, nous avons $\mathbf{x} = x_0$ et $Arw(\mathbf{x}) - \lambda I$ est réduite à $x_0 - \lambda$ qui correspond bien au polynôme (1.4) avec $n = 1$. Supposons que 1.4 soit vraie pour n et montrons qu'elle l'est encore pour $n+1$. Comme $n+1$ est la dimension de $\mathbf{x}$, le déterminant de $Arw(\mathbf{x}) - \lambda I$ sera :

$$\begin{vmatrix} x_0 - \lambda & x_1 & \dots & x_n \\ x_1 & x_0 - \lambda & \mathbf{0}^T & 0 \\ \vdots & \mathbf{0} & \ddots & \mathbf{0} \\ x_n & 0 & \mathbf{0}^T & x_0 - \lambda \end{vmatrix}.$$

En calculant ce déterminant par la règle des cofacteurs sur la dernière colonne, par exemple, et en se servant de l'hypothèse de récurrence, nous arrivons à :

$$\begin{aligned} \det(Arw(\mathbf{x}) - \lambda I) &= (-1)^{n+2}x_n(-1)^{n+1}x_n(x_0 - \lambda)^{n-1} + (x_0 - \lambda)p_n(\mathbf{x}, \lambda) \\ &= -(x_0 - \lambda)^{n-1}x_n^2 + (x_0 - \lambda)^{n-1}(\lambda^2 - 2x_0\lambda + x_0^2 - \|\bar{\mathbf{x}}\|^2) \\ &= (x_0 - \lambda)^{n-1}(\lambda^2 - 2x_0\lambda + x_0^2 - \|\bar{\mathbf{x}}\|^2 - x_n^2) \\ &= p_{n+1}(\mathbf{x}, \lambda), \end{aligned}$$

ce qui montre que (1.4) est vraie $\forall n \geq 1$.

Ainsi, (1.4) nous permet d'affirmer que $\lambda_1 = x_0$, $\lambda_2 = x_0 + \|\bar{\mathbf{x}}\|$ et $\lambda_3 = x_0 - \|\bar{\mathbf{x}}\|$ sont les valeurs propres de $Arw(\mathbf{x})$. La multiplicité de λ_1 vaut $n-2$ et celle de λ_2 et λ_3 vaut 1.
Finalement, nous remarquons que le résultat annoncé est vrai puisque que $Arw(\mathbf{x}) \succcurlyeq 0 (\succ 0)$ si et seulement si $\lambda_{\min}(Arw(\mathbf{x})) \geq 0 (> 0)$, c'est-à-dire si et seulement si $\lambda_3 \geq 0 (> 0)$, qui équivaut à $\mathbf{x} \succcurlyeq_{\mathcal{Q}} 0 (\succ_{\mathcal{Q}} 0)$.

□

Cette proposition témoigne du lien étroit qu'il existe entre SOCP et la programmation semi-définie. En réalité, le cône du second ordre peut être inclus dans celui des matrices semi-définies positives puisque une inégalité du cône du second ordre peut toujours être convertie en une condition de matrice semi-définie positive ou définie positive. Ce qui fait la spécificité de SOCP se situe dans la théorie qui en est à la base comme nous allons le voir par la suite.

Enfin, nous désignons par $\mathrm{bd}\,\mathcal{Q}$ la frontière du cône du second ordre, qui a pour définition :

$$\mathrm{bd}\,\mathcal{Q} = \{\mathbf{x} \in \mathcal{Q} \,|\, x_0 = \|\bar{\mathbf{x}}\| \text{ et } \mathbf{x} \neq \mathbf{0}\}.$$

Nous utilisons également $\mathcal{Q}$, $Arw(\mathbf{x})$, R et $\mathbf{e}$ en accord avec la décomposition en blocs de $\mathbf{x}$; c'est-à-dire si $\mathbf{x} = (\mathbf{x}_1; \ldots; \mathbf{x}_r)$ tel que $\mathbf{x}_i \in I\!R^{n_i}$ pour $i = 1, \ldots, r$, alors

$\mathcal{Q} = \mathcal{Q}_{n_1} \times \ldots \times \mathcal{Q}_{n_r}$
$Arw(\mathbf{x}) = Arw(\mathbf{x}_1) \oplus \ldots \oplus Arw(\mathbf{x}_r)$
$R = R_{n_1} \oplus \ldots \oplus R_{n_r}$
$\mathbf{e} = (\mathbf{e}_{n_1}; \ldots; \mathbf{e}_{n_r})$.

Chapitre 2

Problèmes SOCP

2.1 Formulation de problèmes SOCP

2.1.1 Programmation linéaire

La forme standard d'un problème SOCP ressemble très fortement à celle d'un problème linéaire :

$$\begin{cases} \min & \sum_{i=1}^{k} c_i x_i \\ \text{s.c.} & \sum_{i=1}^{k} x_i \mathbf{a}_i = b \\ & x_i \geq 0 \quad \forall\, i = 1, \ldots, k, \end{cases}$$

où, ici, les variables du problème $x_i \in I\!R$, $i = 1, \ldots, k$, les coefficients de la fonction objectif $c_i \in I\!R$, $i = 1, \ldots, k$, les données $\mathbf{a}_i \in I\!R^m$, $i = 1, \ldots, k$ et $b \in I\!R^m$. En réalité, les inégalités de non-négativité $x_i \geq 0$, $i = 1 \ldots, k$, sont des inégalités du cône du second ordre à une dimension ; c'est-à-dire que ces contraintes sont équivalentes à la contrainte $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_k) \in \mathcal{Q}_1 \times \ldots \times \mathcal{Q}_1$. Nous en concluons que la programmation linéaire est un cas particulier de SOCP.
Aussi, étant donné que $\mathcal{Q}_2 = \{(x_0; x_1) \in I\!R^2 | \, x_0 \geq |x_1|\}$, est une rotation de 45 degrés du quadrant non-négatif, il est clair qu'un problème SOCP dans lequel chaque cône du second ordre est soit $\mathcal{Q}_1$, soit $\mathcal{Q}_2$, peut toujours être transformé en un problème linéaire.

Puisque les cônes du second ordre sont des ensembles convexes, un problème SOCP est un problème de programmation convexe. En outre, si la dimension d'un cône du second ordre est supérieure à 2, ce cône n'est pas polyhédral, et donc en général, la région admissible d'un problème SOCP n'est pas polyhédrale.

2.1.2 Programmation quadratique

Tout comme les problèmes de programmation linéaire, les problèmes de programmation quadratique strictement convexes ont des régions admissibles poly-

hédrales et peuvent être résolus en tant que problème SOCP. Plus précisément, considérons le problème suivant :

$$(PQ)\begin{cases} \min & q(\mathbf{x}) = \mathbf{x}^T Q\mathbf{x} + \mathbf{a}^T\mathbf{x} + \beta \\ \text{s.c.} & A\mathbf{x} = b \\ & \mathbf{x} \geq 0 \end{cases}$$

où $\mathbf{x}, \mathbf{a} \in I\!R^n$, $Q \in I\!R^{n\times n}$ est une matrice symétrique définie positive, $\beta \in I\!R$, $A \in I\!R^{m\times n}$ et $b \in I\!R^m$. Notre but est de réécrire (PQ) sous la forme d'un problème SOCP standard.
Nous y parvenons à condition d'exprimer la fonction objectif $q(\mathbf{x})$ autrement.

$$\begin{aligned} q(\mathbf{x}) &= \mathbf{x}^T Q^{1/2}Q^{1/2}\mathbf{x} + \tfrac{1}{2}\mathbf{a}^T\mathbf{x} + \tfrac{1}{2}\mathbf{a}^T\mathbf{x} + \tfrac{1}{4}\mathbf{a}^T Q^{-1}\mathbf{a} - \tfrac{1}{4}\mathbf{a}^T Q^{-1}\mathbf{a} + \beta \\ &= \langle Q^{1/2}\mathbf{x} + \tfrac{1}{2}Q^{-1/2}\mathbf{a},\, Q^{1/2}\mathbf{x} + \tfrac{1}{2}Q^{-1/2}\mathbf{a}\rangle + \beta - \tfrac{1}{4}\mathbf{a}^T Q^{-1}\mathbf{a} \\ &= \|\bar{\mathbf{u}}\|^2 + \beta - \tfrac{1}{4}\mathbf{a}^T Q^{-1}\mathbf{a} \end{aligned}$$

où $\bar{\mathbf{u}} = Q^{1/2}\mathbf{x} + \frac{1}{2}Q^{-1/2}\mathbf{a}$. Ainsi, le problème (PQ) peut être ramené au problème SOCP équivalent :

$$\begin{cases} \min & u_0 \\ \text{s.c.} & Q^{1/2}\mathbf{x} - \bar{\mathbf{u}} = -\frac{1}{2}Q^{-1/2}\mathbf{a} \\ & A\mathbf{x} = b \\ & \mathbf{x} \geq 0,\ \mathbf{u} = (u_0; \bar{\mathbf{u}}) \succcurlyeq_Q \mathbf{0} \end{cases}$$

qui a la forme standard :

$$(SOCP)\begin{cases} \min & u_0 \\ \text{s.c.} & \begin{pmatrix} Q^{1/2} & \vdots & \mathbf{0} & -I_{n\times n} \\ A & \vdots & \mathbf{0} & O_{n\times n} \end{pmatrix} \begin{pmatrix} \mathbf{x} \\ \cdots \\ \mathbf{u} \end{pmatrix} = \begin{pmatrix} -\frac{1}{2}Q^{-1/2}\mathbf{a} \\ b \end{pmatrix} \\ & \mathbf{x} \geq 0,\ \mathbf{u} = (u_0; \bar{\mathbf{u}}) \succcurlyeq_Q \mathbf{0}. \end{cases}$$

Observons que les variables du problème sont $\mathbf{x} \in I\!R^n$ et $\mathbf{u} \in I\!R^{n+1}$. Bien que ces deux problèmes auront la même solution optimale, la valeur de la fonction objectif de (PQ) différera de la valeur optimale de $(SOCP)$ de $\beta - \frac{1}{4}\mathbf{a}^T Q^{-1}\mathbf{a}$.

Plus généralement, des problèmes quadratiques à contraintes quadratiques convexes peuvent être résolus comme problèmes SOCP. Pour s'en convaincre, considérons le problème général suivant :

$$(PQCQ)\begin{cases} \min & q_0(\mathbf{x}) = \mathbf{x}^T Q_0\mathbf{x} + \mathbf{a}_0^T\mathbf{x} + \beta_0 \\ \text{s.c.} & q_i(\mathbf{x}) = \mathbf{x}^T B_i^T B_i\mathbf{x} + \mathbf{a}_i^T\mathbf{x} + \beta_i \leq 0, \quad i = 1, \ldots, m, \end{cases}$$

où $Q_0 \in I\!R^{n\times n}$ est supposée symétrique définie positive et $B_i \in I\!R^{k_i\times n}$ est supposée de rang k_i, $i = 1, \ldots, m$. Comme nous avons déja montré comment traiter

une fonction objectif quadratique strictement convexe, intéressons-nous uniquement à la transformation des contraintes quadratiques. Chacune des contraintes de (PQCQ) du type

$$q(\mathbf{x}) = \mathbf{x}^T B^T B\mathbf{x} + \mathbf{a}^T\mathbf{x} + \beta \leq 0$$

est équivalente à la contrainte du cône du second ordre $(u_0; \bar{\mathbf{u}}) \succcurlyeq_Q 0$, où

$$\bar{\mathbf{u}} = \begin{pmatrix} B\mathbf{x} \\ \frac{\mathbf{a}^T\mathbf{x}+\beta+1}{2} \end{pmatrix} \quad \text{et} \quad u_0 = \frac{1 - \mathbf{a}^T\mathbf{x} - \beta}{2}.$$

2.1.3 Problèmes de minimisation de normes

SOCP permet de résoudre un nombre important de problèmes de minimisation de normes. En particulier, posons $\bar{\mathbf{v}}_i = A_i\mathbf{x} + b_i \in I\!R^{n_i}$, $i = 1, \ldots, r$. Tous les problèmes de minimisation de normes qui vont suivre vont pouvoir être transformés en problèmes SOCP.

a) Minimiser la somme de r normes :

Le problème $\min \sum_{i=1}^{r} \|\bar{\mathbf{v}}_i\|$ peut être formulé comme :

$$\begin{cases} \min & \sum_{i=1}^{r} v_{i0} \\ \text{s.c.} & A_i\mathbf{x} + b_i = \bar{\mathbf{v}}_i \quad i = 1, \ldots, r \\ & (v_{i0}; \bar{\mathbf{v}}_i) \succcurlyeq_Q 0 \quad i, \ldots, r \end{cases}$$

Les variables du problème sont $\mathbf{x} \in I\!R^n$ et $\mathbf{v}_i = (v_{i0}; \bar{\mathbf{v}}_i) \in I\!R^{m+1}$, $i = 1, \ldots, r$. Le problème de *Fermat-Weber* est un cas particulier de problème de minimisation d'une somme de normes. Il pose le problème de trouver le vecteur $\mathbf{x}$ (symbolisant par exemple un entrepôt) qui minimise la somme des distances de ce point à k vecteurs fixés $\mathbf{d}_i$ (symbolisant par exemple des points de vente). Ce problème est alors formulé de la façon suivante :

$$\min_{\mathbf{x}} \sum_{i=1}^{k} \|\mathbf{d}_i - \mathbf{x}\|$$

où les $\mathbf{d}_i$, $i = 1, \ldots, k$, sont les vecteurs fixés et $\mathbf{x}$ est l'inconnue.

b) Minimiser le maximum de r normes :

Le problème $\min \max_{1 \leq i \leq r} \|\bar{\mathbf{v}}_i\|$ est équivalent au problème SOCP suivant :

$$\begin{cases} \min & t \\ \text{s.c.} & A_i\mathbf{x} + b_i = \bar{\mathbf{v}}_i \quad i = 1, \ldots, r \\ & (t; \bar{\mathbf{v}}_i) \succcurlyeq_Q 0 \quad i, \ldots, r \end{cases}$$

où les variables du problème sont $\mathbf{x} \in I\!R^n$ et $\mathbf{v}_i = (t; \bar{\mathbf{v}}_i) \in I\!R^{m+1}$, $i = 1, \ldots, r$.

c) Minimiser la somme des k plus grandes normes :

Le problème $\min \sum_{i=1}^{k} \|\bar{\mathbf{v}}_{[i]}\|$, où $\|\bar{\mathbf{v}}_{[1]}\|$, $\|\bar{\mathbf{v}}_{[2]}\|, \ldots, \|\bar{\mathbf{v}}_{[r]}\|$ correspondent aux normes $\|\bar{\mathbf{v}}_1\|, \|\bar{\mathbf{v}}_2\|, \ldots, \|\bar{\mathbf{v}}_r\|$ rangées par ordre décroissant, possède la formulation en problème SOCP suivante :

$$\begin{cases} \min & kt + \sum_{i=1}^{r} u_i \\ \text{s.c.} & A_i\mathbf{x} + b_i = \bar{\mathbf{v}}_i \quad i = 1, \ldots, r \\ & \|\bar{\mathbf{v}}_i\| \leq u_i + t, \quad i, \ldots, r \\ & u_i \geq 0 \quad i = 1, \ldots, r \end{cases}$$

en remarquant que

$$\begin{aligned} kt + \sum_{i=1}^{r} u_i &= \underbrace{t + \ldots + t}_{k \text{ fois}} + u_1 + \ldots + u_r \\ &\geq \sum_{i=1}^{k} \|\bar{\mathbf{v}}_{[i]}\|. \end{aligned}$$

2.1.4 Problèmes avec contraintes hyperboliques

Si nous effectuons une rotation de 45 degrés dans le plan (x_0, x_1) dans le sens des $x_0, x_1 \geq 0$, nous obtenons la rotation du cône du second ordre notée $\widehat{\mathcal{Q}}$ et définie par :

$$\widehat{\mathcal{Q}} = \{\mathbf{x} = (x_0; x_1; \hat{\mathbf{x}}) \in I\!R \times I\!R \times I\!R^{n-2} \,|\, 2x_0x_1 \geq \|\hat{\mathbf{x}}\|^2, x_0 \geq 0, x_1 \geq 0\},$$

où par définition $(x_1; \hat{\mathbf{x}}) = \bar{\mathbf{x}}$. On peut se convaincre que $\widehat{\mathcal{Q}}$ est bien une rotation de 45 degrés de $\mathcal{Q}$ dans le plan (x_0, x_1) car, avec $(y_0; y_1; \hat{\mathbf{y}}) \in \mathcal{Q}$ nous avons :

$$\begin{pmatrix} x_0 \\ x_1 \\ \hat{\mathbf{x}} \end{pmatrix} := \begin{pmatrix} \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} & 0 & \ldots & 0 \\ \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} & 0 & \ldots & 0 \\ 0 & 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & 1 \end{pmatrix} \begin{pmatrix} y_0 \\ y_1 \\ \hat{\mathbf{y}} \end{pmatrix} = \begin{pmatrix} \frac{\sqrt{2}}{2}(y_0 - y_1) \\ \frac{\sqrt{2}}{2}(y_0 + y_1) \\ \hat{\mathbf{y}} \end{pmatrix}$$

où, $2x_0x_1 = (y_0 - y_1)(y_0 + y_1) = y_0^2 - y_1^2 \geq \|\bar{\mathbf{y}}\|^2 - y_1^2 = \|\hat{\mathbf{y}}\|^2 = \|\hat{\mathbf{x}}\|^2$. De plus, comme $y_0 \geq \|\bar{\mathbf{y}}\| \geq |y_1|$, nous avons $x_0, x_1 \geq 0$, d'où $\mathbf{x} \in \widehat{\mathcal{Q}}$.

Le cône $\widehat{\mathcal{Q}}$ est très utile pour convertir des problèmes comportant des contraintes hyperboliques en problèmes SOCP. Concrètement, une contrainte de la forme :

$$\mathbf{w}^T\mathbf{w} \leq xy, \text{ où } x \geq 0, y \geq 0, \mathbf{w} \in I\!R^n, x, y \in I\!R$$

est équivalente à une containte du cône du second ordre car

$$\begin{aligned} \mathbf{w}^T\mathbf{w} \leq xy &\iff \mathbf{w}^T\mathbf{w} \leq \tfrac{(x+y)^2}{4} - \tfrac{(x-y)^2}{4} \\ &\iff 4\mathbf{w}^T\mathbf{w} + (x-y)^2 \leq (x+y)^2 \\ &\iff \left\| \begin{pmatrix} 2\mathbf{w} \\ x-y \end{pmatrix} \right\| \leq x+y, \end{aligned}$$

qui est bien une contrainte du cône du second ordre.
Ce type de transformation permet d'étendre de manière significative l'ensemble des problèmes qui peuvent être ramenés à des problèmes SOCP. Illustrons à présent la façon d'obtenir des problèmes SOCP pour plusieurs autres familles de problèmes.

a) Minimiser la moyenne harmonique de fonctions affines positives :

Considérons le problème $\min \sum_{i=1}^{r} \frac{1}{(\mathbf{a}_i^T\mathbf{x} + \beta_i)}$, où $\mathbf{a}_i^T\mathbf{x} + \beta_i > 0$, $i = 1, \ldots, r$.
Posons $v_i = \mathbf{a}_i^T\mathbf{x} + \beta_i$, $i = 1, \ldots, r$. Le problème sera équivalent à celui de minimiser $\sum_{i=1}^{r} u_i$ tq $u_i \geq 1/v_i$ et $u_i \geq 0$ pour $i = 1, \ldots, r$. Autrement dit, nous obtenons le problème équivalent suivant :

$$\begin{cases} \min & \sum_{i=1}^{r} u_i \\ \text{s.c.} & v_i = \mathbf{a}_i^T\mathbf{x} + \beta_i, \quad i = 1, \ldots, r \\ & 1 \leq u_i v_i, \quad i = 1, \ldots, r \\ & u_i \geq 0, \quad i = 1, \ldots, r. \end{cases}$$

b) Approximation logarithmique de Tchebychev :

Le problème de l'approximation logarithmique de Tchebychev s'exprime comme le problème : $\min \max_{1 \leq i \leq r} |\ln(\mathbf{a}_i^T\mathbf{x}) - \ln b_i|$, où b_i, $\mathbf{a}_i^T\mathbf{x} > 0$, $i = 1, \ldots, r$.
La façon de reformuler ce problème en problème SOCP est la suivante :

$$\begin{aligned} |\ln(\mathbf{a}_i^T\mathbf{x}) - \ln b_i| &= \left| \ln\left(\tfrac{\mathbf{a}_i^T\mathbf{x}}{b_i}\right) \right| \\ &= \max\left(\ln\left(\tfrac{\mathbf{a}_i^T\mathbf{x}}{b_i}\right), -\ln\left(\tfrac{\mathbf{a}_i^T\mathbf{x}}{b_i}\right) \right) \\ &= \max\left(\ln\left(\tfrac{\mathbf{a}_i^T\mathbf{x}}{b_i}\right), \ln\left(\tfrac{b_i}{\mathbf{a}_i^T\mathbf{x}}\right) \right) \\ &= \ln \max\left(\tfrac{\mathbf{a}_i^T\mathbf{x}}{b_i}, \tfrac{b_i}{\mathbf{a}_i^T\mathbf{x}} \right) \end{aligned}$$

où la dernière égalité provient du fait que la fonction ln est croissante et continue sur son domaine. Par conséquent, le problème de départ est équivalent au problème :

$$\begin{cases} \min & t \\ \text{s.c.} & 1 \leq (\mathbf{a}_i^T\mathbf{x}/b_i)t \quad i = 1, \ldots, r, \\ & \mathbf{a}_i^T\mathbf{x}/b_i \leq t \quad i = 1, \ldots, r, \\ & t \geq 0. \end{cases}$$

c) Inégalités impliquant la somme de fractions quadratiques/linéaires :

L'inégalité $\sum_{i=1}^{r} \frac{\|A_i\mathbf{x}+b_i\|^2}{\mathbf{a}_i^T\mathbf{x}+\beta_i} \leq t$, où pour chaque i, $A_i\mathbf{x}+b_i = \mathbf{0}$ si $\mathbf{a}_i^T\mathbf{x}+\beta_i = 0$ et $0^2/0 = 0$, peut être représentée par le système d'inégalités du cône du second ordre suivant :

$$\begin{cases} \sum_{i=1}^{r} u_i \leq t, \\ \mathbf{w}_i^T\mathbf{w}_i \leq u_i v_i, \quad i = 1,\ldots,r, \\ \mathbf{w}_i = A_i\mathbf{x}+b_i, \quad i = 1,\ldots,r, \\ v_i = \mathbf{a}_i^T\mathbf{x}+\beta_i \geq 0, \quad i = 1,\ldots,r. \end{cases}$$

d) Problèmes de fractions de matrices :

Considérons les problèmes qui traitent des fractions de matrices de la forme $\mathbf{y}^T A(\mathbf{s})^{-1}\mathbf{y}$, avec $A(\mathbf{s}) = \sum_{i=1}^{k} s_i A_i$, où les $A_i \in I\!R^{n\times n}$, $i = 1,\ldots,k$ sont des matrices symétriques et semi-définies positives dont la somme est définie positive, $\mathbf{y} \in I\!R^n$ et $\mathbf{s} \in I\!R^k_{++}$ (vecteur de dimension k à composantes réelles strictement positives). Nous allons voir que certains de ces problèmes peuvent être formulés en problèmes SOCP. Intéressons-nous, en particulier, à la contrainte d'inégalité suivante :

$$\mathbf{y}^T A(\mathbf{s})^{-1}\mathbf{y} \leq t, \tag{2.1}$$

où $t \in I\!R_+$. Sous l'hypothèse que $\mathbf{s} \in I\!R^k_{++}$ et $\sum_{i=1}^{k} A_i \succ 0$, nous avons que $A(\mathbf{s})$ est définie positive et donc non-singulière. En effet, $\forall\, d \in I\!R^n$, $d \neq 0$,

$$d^T A(\mathbf{s})d = \sum_{i=1}^{k} s_i d^T A_i d \geq \min_{1\leq i\leq k} s_i \; d^T(\sum_{i=1}^{k} A_i)d > 0.$$

Cette propriété étant vérifiée, nous montrons à présent que $(\mathbf{y};\, \mathbf{s};\, t) \in I\!R^n \times I\!R^k_{++} \times I\!R_+$ satisfait (2.1) si et seulement si il existe $\mathbf{w}_i \in I\!R^{r_i}$ et $t_i \in I\!R_+$, $i = 1,\ldots,k$ tels que

$$\begin{array}{l} \sum_{i=1}^{k} D_i^T\mathbf{w}_i = \mathbf{y}, \\ \sum_{i=1}^{k} t_i \leq t, \\ \mathbf{w}_i^T\mathbf{w}_i \leq s_i t_i, \quad i = 1,\ldots,k, \end{array} \tag{2.2}$$

où pour $i = 1,\ldots,k$, r_i =rg(A_i) et $D_i \in I\!R^{r_i\times n}$ tel que $D_i^T D_i = A_i$.
Pour prouver cela, définissons $\mathbf{u}$ par $A(\mathbf{s})\mathbf{u} = \mathbf{y}$ et $\mathbf{w}_i$ par $\mathbf{w}_i = s_i D_i\mathbf{u}$, pour $i = 1,\ldots,k$. Alors, en supposant que $\mathbf{y}$, $\mathbf{s}$ et t satisfont (2.1), nous avons

$$\sum_{i=1}^{k} \frac{\mathbf{w}_i^T\mathbf{w}_i}{s_i} = \sum_{i=1}^{k} \mathbf{u}^T s_i D_i^T D_i\mathbf{u} = \mathbf{u}^T A(\mathbf{s})\mathbf{u} = \mathbf{y}^T A(\mathbf{s})^{-1}\mathbf{y} \leq t$$

et $\sum_{i=1}^{k} D_i^T \mathbf{w}_i = \sum_{i=1}^{k} s_i D_i^T D_i \mathbf{u} = A(\mathbf{s})\mathbf{u} = \mathbf{y}$. Ainsi, en posant $t_i = \frac{\mathbf{w}_i^T \mathbf{w}_i}{s_i}$, $i = 1, \ldots, k$ nous aboutissons à (2.2).
Maintenant, supposons que $(\mathbf{y}; \mathbf{s}; t) \in I\!R^n \times I\!R^k_{++} \times I\!R_+$ et que $((\mathbf{w}_i)_{i=1,\ldots,k}; (t_i)_{i=1,\ldots k})$ est une solution de (2.2). En éliminant les variables t_i dans (2.2), cela revient à supposer que $(\mathbf{y}; \mathbf{s}; t) \in I\!R^n \times I\!R^k_{++} \times I\!R_+$ et que $\mathbf{w}_i \in I\!R^{r_i}$, $i = 1, \ldots, k$ est une solution de

$$\begin{array}{l} \sum_{i=1}^{k} D_i^T \mathbf{w}_i = \mathbf{y}, \\ \sum_{i=1}^{k} \frac{\mathbf{w}_i^T \mathbf{w}_i}{s_i} \leq t. \end{array} \tag{2.3}$$

Considérons, à présent, le problème quadratique

$$\left\{ \begin{array}{ll} \min\limits_{(\omega_1, \ldots, \omega_k)} & \sum_{i=1}^{k} \frac{\omega_i{}^T \omega_i}{s_i} \\ \text{s.c.} & \sum_{i=1}^{k} D_i^T \omega_i = \mathbf{y}, \end{array} \right. \tag{2.4}$$

pour $\mathbf{s} \in I\!R^k_{++}$ et $\mathbf{y}$ satisfaisant (2.3). Ce problème étant convexe, les conditions de KKT seront nécessaires et suffisantes pour la détermination d'une solution optimale $(\omega_1^*, \ldots, \omega_k^*)$. Le langrangien de ce système est

$$L(\omega_1, \ldots, \omega_k; \mathbf{v}) = \sum_{i=1}^{k} \frac{\omega_i{}^T \omega_i}{s_i} + \mathbf{v}^T (\sum_{i=1}^{k} D_i^T \omega_i - \mathbf{y}),$$

où $\mathbf{v} \in I\!R^n$. En posant $\mathbf{u} = -\frac{1}{2}\mathbf{v}$ et en égalant les gradients du lagrangien à 0 pour $i = 1, \ldots, k$ nous concluons par KKT qu'il existe une solution optimale $(\omega_1^*, \ldots, \omega_k^*)$ de (2.4) si et seulement il existe un vecteur $\mathbf{u} \in I\!R^n$ tel que

$$\omega_i^* = s_i D_i \mathbf{u}, \; i = 1, \ldots, k \quad \text{et} \quad \sum_{i=1}^{k} D_i^T \omega_i^* = \mathbf{y}.$$

Mais puisque les $\mathbf{w}_i$ vérifient les équations de (2.3), cela implique que

$$\mathbf{u}^T A(\mathbf{s})\mathbf{u} = \sum_{i=1}^{k} s_i \mathbf{u}^T D_i^T D_i \mathbf{u} = \sum_{i=1}^{k} \frac{\omega_i^{*T} \omega_i^*}{s_i} \leq \sum_{i=1}^{k} \frac{\mathbf{w}_i^T \mathbf{w}_i}{s_i} \leq t.$$

De plus,

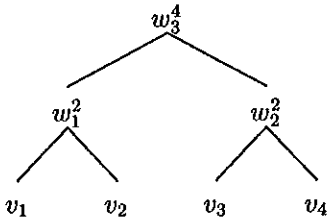
$$A(\mathbf{s})\mathbf{u} = \sum_{i=1}^{k} s_i D_i^T D_i \mathbf{u} = \sum_{i=1}^{k} D_i^T \omega_i^* = \mathbf{y}$$

qui nous permet d'écrire $\mathbf{y}^T A(\mathbf{s})^{-1} \mathbf{y} = \mathbf{u}^T A(\mathbf{s})\mathbf{u} \leq t$. Par conséquent, $(\mathbf{y}; \mathbf{s}; t)$ est une solution de (2.1).

e) Maximiser la moyenne géométrique de fonctions affines non-négatives :

Pour illustrer la manière de transformer le problème $\max \prod_{i=1}^{r} (\mathbf{a}_i^T \mathbf{x} + \beta_i)^{1/r}$ en un problème comportant des contraintes hyperboliques, nous montrons comment procéder dans le cas $r = 4$, afin d'avoir les idées claires.

Posons tout d'abord $v_i = \mathbf{a}_i^T\mathbf{x} + \beta_i \geq 0$, $i = 1, \ldots, 4$ ainsi que w_1, $w_2 \in I\!R_+$ et $w_3 \in I\!R$. Considérons ensuite l'arbre binaire suivant :



où la relation

représente l'inégalité $x \leq yz$. A partir de cet arbre, il est clair qu'en maximisant la racine (modulo puissance 1/4) nous allons maximiser $\prod_{i=1}^4 v_i^{1/4}$. Ainsi, nous obtenons le problème à contraintes hyperboliques suivant :

$$\begin{cases} \max & w_3 \\ \text{s.c.} & v_i = \mathbf{a}_i^T\mathbf{x} + \beta_i \geq 0,\ i = 1, \ldots, 4 \\ & w_1^2 \leq v_1 v_2, \quad w_2^2 \leq v_3 v_4, \quad w_3^2 \leq w_1 w_2, \\ & w_1 \geq 0, \quad w_2 \geq 0. \end{cases}$$

En formulant le problème de la moyenne géométrique ci-dessus comme un problème SOCP, nous avons utilisé le fait qu'une inégalité de la forme :

$$t^{2^k} \leq s_1 s_2 \ldots s_{2^k} \tag{2.5}$$

pour $t \in I\!R$, et $s_1 \geq 0, \ldots, s_{2^k} \geq 0$ peut être exprimée par 2^{k-1} inégalités de la forme $w_i^2 \leq u_i v_i$, où toutes les nouvelles variables qui sont introduites doivent être non-négatives.

f) Problèmes impliquant des paires de formes quadratiques :

En guise d'illustration de tels problèmes, considérons ici le problème de trouver la sphère la moins volumineuse $S = \{\mathbf{x} \in I\!R^n \,|\, \|\mathbf{x} - \mathbf{a}\| \leq \rho\}$, qui contient des ellipsoïdes $\mathcal{E}_1, \ldots, \mathcal{E}_k$ données, où

$$\mathcal{E}_i = \{\mathbf{x} \in I\!R^n \,|\, \mathbf{x}^T A_i \mathbf{x} + 2\mathbf{b}_i^T\mathbf{x} + c_i \leq 0\}, \quad \text{pour } i = 1, \ldots, k$$

tel que $\forall i,\ A_i = A_i^T$. Le fait que chaque ellispoïde $\mathcal{E}_i$ doit être contenue dans la sphère se traduit par k conditions mettant chacune en relation une paire de fonctions quadratiques. Plus précisément, nous avons les k conditions suivantes :

$$\begin{cases} \mathbf{x}A_1^T\mathbf{x} + 2\mathbf{b}_1^T\mathbf{x} + c_1 \leq 0 & \Longrightarrow & \|\mathbf{x} - \mathbf{a}\| \leq \rho \\ \qquad\vdots & \vdots & \vdots \\ \mathbf{x}A_k^T\mathbf{x} + 2\mathbf{b}_k^T\mathbf{x} + c_k \leq 0 & \Longrightarrow & \|\mathbf{x} - \mathbf{a}\| \leq \rho \end{cases} \tag{2.6}$$

Dans ce genre cas, où une fonction quadratique F_0 est contrainte à être non-négative lorsque une autre fonction quadratique F_1 l'est, il est pratique d'utiliser

la $\mathcal{S}$-procédure. Grâce à la $\mathcal{S}$-procédure, il nous sera possible de remplacer le système de contraintes (2.6) par des inégalités matricielles linéaires en les données définissant les fonctions quadratiques. L'énoncé de la $\mathcal{S}$-procédure (dont la preuve figure dans les annexes) pour une paire de fonctions quadratiques est le suivant :

Proposition 2.1.1 ($\mathcal{S}$-procédure)
Soient $F_0(x) = x^T Q_0 x + 2s_0^T x + r_0$ et $F_1(x) = x^T Q_1 x + 2s_1^T x + r_1$ deux fonctions quadratiques définies sur $\mathbb{R}^n$ et à valeurs dans $\mathbb{R}$ telles que Q_0 et Q_1 soient symétriques et telles qu'il existe $\tilde{x} \in \mathbb{R}^n$ vérifiant $F_1(\tilde{x}) > 0$. Alors, les deux affirmations suivantes sont équivalentes :

1. *$F_0(x) \geq 0 \quad \forall x \in \mathbb{R}^n$ tel que $F_1(x) \geq 0$.*
2. *Il existe $\tau \geq 0$ tel que*

$$\begin{bmatrix} Q_0 & s_0 \\ s_0^T & r_0 \end{bmatrix} - \tau \begin{bmatrix} Q_1 & s_1 \\ s_1^T & r_1 \end{bmatrix} \succcurlyeq 0.$$

Notons que

$$\begin{aligned} \|\mathbf{x} - \mathbf{a}\| \leq \rho \quad &\Longleftrightarrow \quad \|\mathbf{x} - \mathbf{a}\|^2 \leq \rho^2 \\ &\Longleftrightarrow \quad \mathbf{x}^T\mathbf{x} - 2\mathbf{a}^T\mathbf{x} + \|\mathbf{a}\|^2 - \rho^2 \leq 0. \end{aligned}$$

Aussi, en vue d'appliquer la $\mathcal{S}$-procédure avec les notations de la proposition (2.1.1), effectuons les affectations suivantes :

$$\begin{array}{ll} Q_0 \hookleftarrow -I & Q_i \hookleftarrow -A_i \quad i = 1, \ldots, k, \\ s_0 \hookleftarrow \mathbf{a} & s_i \hookleftarrow -\mathbf{b}_i \quad i = 1, \ldots, k, \\ r_0 \hookleftarrow \rho^2 - \|\mathbf{a}\|^2 & r_i \hookleftarrow -c_i \quad i = 1, \ldots, k. \end{array}$$

En appliquant la $\mathcal{S}$-procedure à chaque condition figurant dans (2.6) nous obtenons :

S contient $\mathcal{E}_1, \ldots, \mathcal{E}_k$ si et seulement si il existe des scalaires non-négatifs $\tau_1, \ldots, \tau_k$ tels que

$$M_i := \begin{pmatrix} \tau_i A_i - I & \tau_i \mathbf{b}_i + \mathbf{a} \\ \tau_i \mathbf{b}_i^T + \mathbf{a}^T & \tau_i c_i + \rho^2 - \mathbf{a}^T\mathbf{a} \end{pmatrix} \succcurlyeq 0, \qquad \text{pour } i = 1, \ldots, k.$$

Considérons ensuite la factorisation de Jordan de A_i, $A_i = Q_i \Lambda_i Q_i^T$ où Q_i est orthogonale et $\Lambda_i = \mathrm{diag}(\lambda_{i1}, \ldots, \lambda_{in})$. Posons $t = \rho^2$ et $\mathbf{v}_i = Q_i^T(\tau_i \mathbf{b}_i + \mathbf{a})$. Alors,

$$\bar{M}_i := \begin{pmatrix} Q_i^T & \mathbf{0} \\ \mathbf{0}^T & 1 \end{pmatrix} M_i \begin{pmatrix} Q_i & \mathbf{0} \\ \mathbf{0}^T & 1 \end{pmatrix} = \begin{pmatrix} \tau_i \Lambda_i - I & \mathbf{v}_i \\ \mathbf{v}_i^T & \tau_i c_i + t - \mathbf{a}^T\mathbf{a} \end{pmatrix} \succcurlyeq 0,$$

pour $i = 1, \ldots, k$, et $M_i \succcurlyeq 0 \Leftrightarrow \bar{M}_i \succcurlyeq 0$. Or, cette condition de semi-définie positivité sur $\bar{M}_i$ survient si et seulement si $\tau_i \geq \frac{1}{\lambda_{\min}(A_i)}$, i.e., $\tau_i \lambda_{ij} - 1 \geq 0$

pour tout i, j, si $v_{ij} = 0$ lorsque $\tau_i \lambda_{ij} - 1 = 0$ et si le complément de Schur des colonnes et des lignes de $\bar{M}_i$ qui ne sont pas nulles est positif ou nul, i.e.,

$$\tau_i c_i + t - \mathbf{a}^T \mathbf{a} - \sum_{j \; tq \; \tau_i \lambda_{ij} > 1} \frac{v_{ij}^2}{(\tau_i \lambda_{ij} - 1)} \geq 0 \tag{2.7}$$

Si nous définissons $\mathbf{s}_i = (s_{i1}; \ldots; s_{in})$, où $s_{ij} = \frac{v_{ij}^2}{\tau_i \lambda_{ij} - 1}$, pour tout j tq $\tau_i \lambda_{ij} > 1$ et $s_{ij} = 0$, sinon, alors (2.7) est équivalent à

$$t \geq \mathbf{a}^T \mathbf{a} - \tau_i c_i + 1^T \mathbf{s}_i.$$

Puisque nous minimisons t, nous pouvons remplacer l'égalité définissant les s_{ij} par la contrainte $v_{ij}^2 \leq s_{ij}(\tau_i \lambda_{ij} - 1)$, $j = 1, \ldots, n$, $i = 1, \ldots, k$. Combinant tous les résultats obtenus, nous aboutissons à la formulation suivante qui utilise uniquement des contraintes linéaires et hyperboliques :

$$\left\{ \begin{array}{ll} \min & t \\ \text{s.c.} & \mathbf{v}_i = Q_i^T(\tau_i \mathbf{b}_i + \mathbf{a}), \quad i = 1, \ldots, k \\ & v_{ij}^2 \leq s_{ij}(\tau_i \lambda_{ij} - 1), \quad i = 1, \ldots, k, \text{ et } j = 1, \ldots, n \\ & \mathbf{a}^T \mathbf{a} \leq \sigma \\ & \sigma \leq t + \tau_i c_i - 1^T \mathbf{s}_i, \quad i = 1, \ldots, k \\ & \tau_i \geq \frac{1}{\lambda_{\min}(A_i)}, \quad i = 1, \ldots, k. \end{array} \right.$$

Cet exemple illustre le fait que bien qu'il soit possible d'exprimer un problème comme un problème SOCP, cette tâche est loin d'être triviale. De plus, dans tous les cas de figure envisagés nous avons pu observer qu'il était parfois commode d'introduire un grand nombre de variables supplémentaires par rapport au nombre de variables au départ. Cependant, cette augmentation du nombre de variables n'est pas forcément un inconvénient d'un point de vue de la rapidité algorithmique puisque, comme nous le verrons, la complexité des algorithmes de points intérieurs dépend du nombre d'inégalités du cône du second ordre et non, directement, du nombre de variables.

2.2 Problèmes robustes

La détermination de solutions robustes pour des problèmes d'optimisation a été pendant quelque temps un sujet à la base de nombreuses recherches dans le domaine de la théorie du contrôle. Récemment, l'idée de robustesse a été introduite dans les domaines de la programmation mathématique et des moindres carrés. Dans cette section, nous allons montrer que les équivalents robustes du problème des moindres carrés et d'un programme linéaire peuvent être tous les deux formulés comme des problèmes SOCP.

2.2.1 Programmation linéaire robuste

Considérons un programme linéaire de la forme

$$\begin{cases} \min & \mathbf{c}^T\mathbf{x} \\ \text{s.c.} & \mathbf{a}_i^T\mathbf{x} \leq b_i, \quad i = 1, \ldots, m, \end{cases} \tag{2.8}$$

où les données $\mathbf{c}$, $\mathbf{a}_i \in I\!R^n$ et $b_i \in I\!R$ sont connues à une certaine incertitude près ou sont soumises à des variations. Pour simplifier la présentation, nous supposerons que $\mathbf{c}$ et b_i sont fixés, et que les $\mathbf{a}_i$ se situent dans des ellipsoïdes $\mathcal{E}_i$ donnés par

$$\mathbf{a}_i \in \mathcal{E}_i \stackrel{\text{déf}}{=} \{\bar{\mathbf{a}}_i + P_i\mathbf{u} \,|\, \|\mathbf{u}\| \leq 1\}$$

où $P_i = P_i^T \succcurlyeq 0$. Connaissant les domaines d'incertitude sur les $\mathbf{a}_i$, nous imposerons que les contraintes soient satisfaites pour toutes les valeurs possibles des paramètres $\mathbf{a}_i$, ce qui nous conduit au programme linéaire robuste suivant :

$$\begin{cases} \min & \mathbf{c}^T\mathbf{x} \\ \text{s.c.} & \mathbf{a}_i^T\mathbf{x} \leq b_i, \quad \forall\, \mathbf{a}_i \in \mathcal{E}_i, \quad i = 1, \ldots, m. \end{cases} \tag{2.9}$$

La contrainte linéaire robuste $\mathbf{a}_i^T\mathbf{x} \leq b_i, \ \forall\, \mathbf{a}_i \in \mathcal{E}_i$ peut être exprimée comme

$$\bar{\mathbf{a}}_i{}^T\mathbf{x} + \|P_i\mathbf{x}\| \leq b_i.$$

En effet,

$$\begin{array}{lll} \mathbf{a}_i^T\mathbf{x} \leq b_i, \quad \forall\, \mathbf{a}_i \in \mathcal{E}_i & \Longleftrightarrow & \max\{\mathbf{a}_i^T\mathbf{x} \,|\, \mathbf{a}_i \in \mathcal{E}_i\} \leq b_i \\ & \Longleftrightarrow & \max\{\bar{\mathbf{a}}_i{}^T\mathbf{x} + \mathbf{u}^T P_i\mathbf{x} \,|\, \|\mathbf{u}\| \leq 1\} \leq b_i \\ & \Longleftrightarrow & \bar{\mathbf{a}}_i{}^T\mathbf{x} + \|P_i\mathbf{x}\| \leq b_i, \end{array}$$

où, dans la dernière équivalence, le maximum est obtenu avec $\mathbf{u} = \frac{P_i\mathbf{x}}{\|P_i\mathbf{x}\|}$ si $P_i\mathbf{x} \neq \mathbf{0}$, et pour n'importe quel $\mathbf{u}$ tel que $\|\mathbf{u}\| = 1$, si $P_i\mathbf{x} = \mathbf{0}$. Donc, le programme robuste (2.9) est donné par le problème SOCP équivalent suivant :

$$\begin{cases} \min & \mathbf{c}^T\mathbf{x} \\ \text{s.c.} & \bar{\mathbf{a}}_i{}^T\mathbf{x} + \|P_i\mathbf{x}\| \leq b_i, \quad i = 1, \ldots, m, \end{cases}$$

ou encore, dans une forme plus standard,

$$\begin{cases} \min & \mathbf{c}^T\mathbf{x} \\ \text{s.c.} & \bar{\mathbf{a}}_i{}^T\mathbf{x} + t_i = b_i, \quad i = 1, \ldots, m, \\ & P_i\mathbf{x} = d_i, \quad i = 1, \ldots, m, \\ & \|d_i\| \leq t_i, \quad i = 1, \ldots, m. \end{cases}$$

Illustration dans un cadre statistique

La programmation linéaire robuste peut avoir des applications dans un cadre statistique. Nous supposerons ici que les paramètres $\mathbf{a}_i$ sont des vecteurs aléatoires gaussiens indépendants de moyenne $\bar{\mathbf{a}}_i$ et dont la matrice de covariance est Σ_i (matrice symétrique définie positive). Un vecteur aléatoire $\mathbf{a}_i$ a pour fonction de densité :

$$f_{\mathbf{a}_i}(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^n |\Sigma_i|}} e^{-\frac{1}{2}(\mathbf{x}-\bar{\mathbf{a}}_i)^T \Sigma_i{}^{-1}(\mathbf{x}-\bar{\mathbf{a}}_i)}, \quad \forall \mathbf{x} \in I\!R^n.$$

Nous aimerions exprimer le fait que dans le programme linéaire (2.8) chaque contrainte $\mathbf{a}_i^T\mathbf{x} \leq b_i$ soit satisfaite avec une probabilité supérieure ou égale à un paramètre $\eta \geq 0.5$, i.e.,

$$\text{Prob}(\mathbf{a}_i^T\mathbf{x} \leq b_i) \geq \eta. \tag{2.10}$$

Nous allons montrer que cette contrainte de probabilité peut être convertie en une contrainte du cône du second ordre.
En posant $u = \mathbf{a}_i^T\mathbf{x}$ (variable aléatoire normale unidimensionnelle) avec $\bar{u}$ désignant sa moyenne et σ sa variance, cette contrainte peut s'écrire :

$$\text{Prob}\left(\frac{u - \bar{u}}{\sqrt{\sigma}} \leq \frac{b_i - \bar{u}}{\sqrt{\sigma}}\right) \geq \eta. \tag{2.11}$$

Puisque $(u - \bar{u})/\sqrt{\sigma}$ est une variable aléatoire normale unidimensionnelle de moyenne 0 et de variance unitaire, la probabilité figurant dans (2.11) est simplement $\Phi((b_i - \bar{u})/\sqrt{\sigma})$, où

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-t^2/2} dt, \quad \forall z \in I\!R$$

est la fonction de répartition d'une variable aléatoire normale de moyenne 0 et de variance unitaire. Ainsi, les contraintes (2.11), et donc (2.10), sont équivalentes à

$$\Phi\left(\frac{b_i - \bar{u}}{\sqrt{\sigma}}\right) \geq \eta$$

ou encore, après application de Φ^{-1} qui est croissante,

$$\bar{u} + \Phi^{-1}(\eta)\sqrt{\sigma} \leq b_i.$$

De plus, des règles de probabilités nous permettent d'obtenir $\bar{u} = \bar{\mathbf{a}}_i{}^T\mathbf{x}$ ainsi que $\sigma = \mathbf{x}^T\Sigma_i\mathbf{x} = \mathbf{x}^T\Sigma_i^{1/2}\Sigma_i^{1/2}\mathbf{x} = \|\Sigma_i^{1/2}\mathbf{x}\|^2$ afin d'aboutir à

$$\bar{\mathbf{a}}_i{}^T\mathbf{x} + \Phi^{-1}(\eta)\|\Sigma_i^{1/2}\mathbf{x}\| \leq b_i.$$

Enfin, puisque $\eta \geq 1/2$ par hypothèse, nous avons $\Phi^{-1}(\eta) \geq \Phi^{-1}(1/2) = 0$ (car $\Phi(0) = 1/2$) ce qui nous assure que la dernière contrainte obtenue est une contrainte du cône du second ordre.

En résumé, le problème

$$\begin{cases} \min & \mathbf{c}^T\mathbf{x} \\ \text{s.c.} & \text{Prob}(\mathbf{a}_i^T\mathbf{x} \leq b_i) \geq \eta, \quad i = 1, \ldots, m \text{ avec } \eta \geq 1/2 \end{cases}$$

est équivalent au problème SOCP suivant

$$\begin{cases} \min & \mathbf{c}^T\mathbf{x} \\ \text{s.c.} & \bar{\mathbf{a}}_i{}^T\mathbf{x} + \Phi^{-1}(\eta)\|\Sigma_i^{1/2}\mathbf{x}\| \leq b_i \quad i = 1, \ldots, m. \end{cases}$$

où la seule inconnue est $\mathbf{x}$.

2.2.2 Moindres carrés robustes

Considérons à présent le système d'équations surdéterminé

$$A\mathbf{x} \approx b, \tag{2.12}$$

où $A \in I\!R^{m\times n}(m > n)$ et $b \in I\!R^m$ sont soumis à des erreurs δA et δb inconnues mais bornées ; nous supposerons que

$$\|(\delta A,\, \delta b)\|_F \leq \rho, \tag{2.13}$$

où $\|B\|_F$ représente la norme de Frobénius d'une matrice B de dimension quelconque. Nous définissons la solution robuste des moindres carrés comme la solution $\hat{\mathbf{x}} \in I\!R^n$ qui minimise le résidu dans le pire des cas, i.e., $\hat{\mathbf{x}}$ est solution de

$$\min_{\mathbf{x}} \max \left\{ \|(A + \delta A)\mathbf{x} - (b + \delta b)\| \mid \|(\delta A,\, \delta b)\|_F \leq \rho \right\}. \tag{2.14}$$

Définissons pour un vecteur $\mathbf{x}$ donné

$$r(A,\, b,\, \mathbf{x}) = \max \left\{ \|(A + \delta A)\mathbf{x} - (b + \delta b)\| \mid \|(\delta A,\, \delta b)\|_F \leq \rho \right\}.$$

Par l'inégalité triangulaire, nous avons

$$\|(A + \delta A)\mathbf{x} - (b + \delta b)\| \leq \|A\mathbf{x} - b\| + \left\| (\delta A,\, -\delta b) \begin{pmatrix} \mathbf{x} \\ 1 \end{pmatrix} \right\|. \tag{2.15}$$

Sachant que pour toute matrice $B \in \mathbb{R}^{m\times n}$ et tout vecteur $x \in \mathbb{R}^n$, $\|Bx\| \leq \|B\|_F\|x\|$, il suit que

$$\left\|(\delta A,\ -\delta b)\begin{pmatrix} \mathrm{x} \\ 1 \end{pmatrix}\right\| \leq \|(\delta A,\ -\delta b)\|_F \left\|\begin{pmatrix} \mathrm{x} \\ 1 \end{pmatrix}\right\| \leq \rho \left\|\begin{pmatrix} \mathrm{x} \\ 1 \end{pmatrix}\right\|.$$

Mais en effectuant le choix $(\delta A,\ -\delta b) = u\mathrm{v}^T$, où

$$u = \begin{cases} \rho\frac{A\mathrm{x}-b}{\|A\mathrm{x}-b\|}, \text{ si } A\mathrm{x}-b \neq 0 \\ \text{vecteur de } \mathbb{R}^m \text{ de norme } \rho, \text{ sinon} \end{cases} \qquad \text{et} \quad \mathrm{v} = \frac{\begin{pmatrix} \mathrm{x} \\ 1 \end{pmatrix}}{\left\|\begin{pmatrix} \mathrm{x} \\ 1 \end{pmatrix}\right\|},$$

tq

$$\begin{aligned} \|(\delta A,\ \delta b)\|_F &= \|(\delta A,\ -\delta b)\|_F = \|u\mathrm{v}^T\|_F \\ &= \sqrt{\textstyle\sum_{i=1}^{m}\sum_{j=1}^{n+1}(u_i\mathrm{v}_j)^2} = \sqrt{\textstyle\sum_{i=1}^{m} u_i^2 \sum_{j=1}^{n+1}\mathrm{v}_j^2} \\ &= \|u\|\|\mathrm{v}\| = \rho, \end{aligned}$$

$A\mathrm{x} - b$ sera multiple de $(\delta A,\ -\delta b)\begin{pmatrix} \mathrm{x} \\ 1 \end{pmatrix}$, car

$$(\delta A,\ -\delta b)\begin{pmatrix} \mathrm{x} \\ 1 \end{pmatrix} = u\mathrm{v}^T\begin{pmatrix} \mathrm{x} \\ 1 \end{pmatrix} = u\left\|\begin{pmatrix} \mathrm{x} \\ 1 \end{pmatrix}\right\| = \rho\frac{A\mathrm{x}-b}{\|A\mathrm{x}-b\|}\left\|\begin{pmatrix} \mathrm{x} \\ 1 \end{pmatrix}\right\|,$$

si $A\mathrm{x}-b \neq 0$. Si $A\mathrm{x}-b = 0$, c'est évident. Cette remarque nous permet d'obtenir l'égalité dans (2.15). D'où,

$$r(A,\ b,\ \mathrm{x}) = \|A\mathrm{x}-b\| + \left\|(\delta A,\ -\delta b)\begin{pmatrix} \mathrm{x} \\ 1 \end{pmatrix}\right\| = \|A\mathrm{x}-b\| + \rho\left\|\begin{pmatrix} \mathrm{x} \\ 1 \end{pmatrix}\right\|.$$

Par conséquent, le problème robuste des moindres carrés est un problème de minimisation d'une somme de normes et peut donc être formulé comme un problème SOCP. Cette formulation est la suivante :

$$\begin{cases} \min & \lambda + \rho\tau \\ \text{s.c.} & \|A\mathrm{x}-b\| \leq \lambda \\ & \left\|\begin{pmatrix} \mathrm{x} \\ 1 \end{pmatrix}\right\| \leq \tau. \end{cases}$$

Chapitre 3

Algèbre du cône du second ordre

Nous allons voir dans ce chapitre qu'il existe une algèbre particulière associée à la programmation du cône du second ordre, ou plus précisement au cône du second ordre. La compréhension de cette algèbre nous permettra d'aborder la plupart des propriétés liées aux problèmes SOCP, comme les notions de dualité, complémentarité, non-dégénérescence, complémentarité stricte, et aboutira au développement d'algorithmes de points intérieurs.

3.1 Algèbre de Jordan

3.1.1 Notion d'algèbre

Dans les définitions qui vont suivre nous allons considérer un espace vectoriel V défini sur $\mathbb{R}$.

Définition 3.1.1 *Le couple $(V, *)$ est une algèbre si V est un espace vectoriel sur $\mathbb{R}$ et $*$ un opérateur tels que $\forall\, \mathbf{x}, \mathbf{y}, \mathbf{z} \in V$ et $\forall\, \alpha, \beta \in \mathbb{R}$,*

$$\begin{aligned} &\mathbf{x} * \mathbf{y} \in V \quad \textit{(fermeture)} \\ &\mathbf{x} * (\alpha\mathbf{y} + \beta\mathbf{z}) = \alpha(\mathbf{x} * \mathbf{y}) + \beta(\mathbf{x} * \mathbf{z}) \quad \textit{(distributivité)} \\ &(\alpha\mathbf{y} + \beta\mathbf{z}) * \mathbf{x} = \alpha(\mathbf{y} * \mathbf{x}) + \beta(\mathbf{z} * \mathbf{x}) \end{aligned}$$

Notons que la règle de distributivité implique que l'opération binaire $\mathbf{x}*\mathbf{y}$ est une fonction bilinéaire en $\mathbf{x}$ et $\mathbf{y}$. En d'autres termes, il doit exister des matrices

carrées $Q_1, Q_2, \ldots, Q_n$ telles que

$$\forall \mathbf{x}, \mathbf{y}, \quad \mathbf{x} * \mathbf{y} = \mathbf{z} = \begin{pmatrix} \mathbf{x}^T Q_1 \mathbf{y} \\ \mathbf{x}^T Q_2 \mathbf{y} \\ \vdots \\ \mathbf{x}^T Q_n \mathbf{y} \end{pmatrix} \text{ (bilinéarité).}$$

Par conséquent, il est clair que la connaissance des matrices $Q_1, \ldots, Q_n$ détermine la loi de multiplication, qui à son tour, détermine l'algèbre. De plus, pour un $\mathbf{x} \in V$ fixé nous avons : $\forall \mathbf{y} \in V$, $\mathbf{x} * \mathbf{y} = L(\mathbf{x})\mathbf{y}$, où $L(\mathbf{x})$ dépend linéairement de $\mathbf{x}$. Donc, $L(\cdot)$ détermine aussi l'algèbre $(V, *)$.
Très souvent, par abus de notation et lorsque le contexte le permettra, nous parlerons de l'algèbre V sans spécifier la loi $*$.

Définition 3.1.2 *Soit une algèbre* $(V, *)$. *Si il existe un élément* $\mathbf{e} \in V$ *tel que* $\forall \mathbf{x} \in V$

$$\mathbf{e} * \mathbf{x} = \mathbf{x} * \mathbf{e} = \mathbf{x},$$

alors $\mathbf{e}$ *est l'élément identité de* $(V, *)$.

Si une algèbre possède un tel élément, cet élément doit être unique, puisque s'il en existait deux, disons $\mathbf{e}_1$ et $\mathbf{e}_2$, nous aurions

$$\mathbf{e}_1 = \mathbf{e}_1 * \mathbf{e}_2 = \mathbf{e}_2.$$

Définition 3.1.3 *Une algèbre* $(V, *)$ *est dite associative si* $\forall \mathbf{x}, \mathbf{y}, \mathbf{z} \in V$,

$$(\mathbf{x} * \mathbf{y}) * \mathbf{z} = \mathbf{x} * (\mathbf{y} * \mathbf{z})$$

Exemples :

1) Soit $\mathcal{M}_n$ l'ensemble des matrices carrées de dimension n. Le couple $(\mathcal{M}_n, .)$, où . désigne la multiplication matricielle classique, constitue une algèbre associative (mais pas commutative). De plus, en définissant

$$\mathbf{vec}(X) = (x_{11}, x_{21}, \ldots, x_{n1}, \ldots, x_{1n}, x_{2n}, \ldots, x_{nn})^T \in I\!R^{n^2}$$

et

$$A \otimes B = \begin{pmatrix} a_{11}B & a_{12}B & \ldots & a_{1n}B \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1}B & a_{n2}B & \ldots & a_{nn}B \end{pmatrix}, \quad \text{où } A, B \in \mathcal{M}_n,$$

nous avons, $\forall X \in \mathcal{M}_n$, $L(X) = I \otimes X$. En effet, soient A, $B \in \mathcal{M}_n$, nous avons

$$\begin{aligned}\mathbf{vec}(AB) &= \begin{pmatrix} (AB)_1 \\ \vdots \\ (AB)_n \end{pmatrix} = \begin{pmatrix} Ab_1 \\ \vdots \\ Ab_n \end{pmatrix} \\ &= \begin{pmatrix} A & 0 & \dots & 0 \\ 0 & A & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \dots & 0 & A \end{pmatrix} \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix} = (I \otimes A)\mathbf{vec}(B)\end{aligned}$$

où $(AB)_i$ et b_i désignent respectivement la i-ème colonne de AB et la i-ème colonne de B.

2) Soit $(\mathcal{M}_n, \circ)$, où ici $\circ$ est tel que pour A, $B \in \mathcal{M}_n$, $A \circ B \overset{\text{déf}}{=} \frac{AB+BA}{2}$. Alors $(\mathcal{M}_n, \circ)$ désigne une algèbre commutative mais pas associative. De plus, $\forall X \in \mathcal{M}_n$, $L(X) = \frac{I \otimes X + X^T \otimes I}{2}$, car nous avons $\forall X, Y \in \mathcal{M}_n$,

$$\mathbf{vec}(X \circ Y) = \mathbf{vec}\left(\frac{XY + YX}{2}\right)$$

$$= \frac{1}{2}\mathbf{vec}(XY) + \frac{1}{2}\mathbf{vec}(YX) = \frac{1}{2}\begin{pmatrix} (XY)_1 \\ \vdots \\ (XY)_n \end{pmatrix} + \frac{1}{2}\begin{pmatrix} (YX)_1 \\ \vdots \\ (YX)_n \end{pmatrix}$$

$$= \frac{1}{2}\begin{pmatrix} X & 0 & \dots & 0 \\ 0 & X & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \dots & 0 & X \end{pmatrix}\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} + \frac{1}{2}\begin{pmatrix} x_{11}I & x_{21}I & \dots & x_{n1}I \\ \vdots & \vdots & \ddots & \vdots \\ x_{1n}I & x_{2n}I & \dots & x_{nn}I \end{pmatrix}\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

$$= \frac{1}{2}(I \otimes X)\mathbf{vec}(Y) + \frac{1}{2}(X^T \otimes I)\mathbf{vec}(Y) = L(X)\mathbf{vec}(Y).$$

3)(Algèbre des formes quadratiques)

Définissons $\forall \mathbf{x} = (x_0; \bar{\mathbf{x}})$, $\mathbf{y} = (y_0; \bar{\mathbf{y}}) \in \mathbb{R}^{n+1}$

$$\mathbf{x} \circ \mathbf{y} = \begin{pmatrix} x_0 y_0 + \bar{\mathbf{x}}^T B \bar{\mathbf{y}} \\ x_0 \bar{\mathbf{y}} + y_0 \bar{\mathbf{x}} \end{pmatrix} \in \mathbb{R}^{n+1},$$

où B est une matrice symétrique $n \times n$. Le couple $(\mathbb{R}^{n+1}, \circ)$ est une algèbre commutative, non-associative et dont l'élément identité est $\mathbf{e} = (1; \mathbf{0})$. Cette algèbre est caractérisée par la transformation linéaire $L(\cdot)$ déterminée comme suit

$$\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^{n+1}, \quad \mathbf{x} \circ \mathbf{y} = \begin{pmatrix} x_0 y_0 + (\bar{\mathbf{x}}^T B)\bar{\mathbf{y}} \\ y_0 \bar{\mathbf{x}} + x_0 I \bar{\mathbf{y}} \end{pmatrix} = \begin{pmatrix} x_0 & \bar{\mathbf{x}}^T B \\ \bar{\mathbf{x}} & x_0 I \end{pmatrix}\begin{pmatrix} y_0 \\ \bar{\mathbf{y}} \end{pmatrix}.$$

D'où,

$$L(\mathbf{x}) = \begin{pmatrix} x_0 & \bar{\mathbf{x}}^T B \\ \bar{\mathbf{x}} & x_0 I \end{pmatrix} \in I\!R^{(n+1)\times(n+1)}.$$

Remarquons que lorsqu'une algèbre $(V, *)$ est associative et déterminée par une matrice $L(\cdot)$, nous avons $\forall \mathbf{x}, \mathbf{y}, \mathbf{z} \in V$,

$$(\mathbf{x} * \mathbf{y}) * \mathbf{z} = \mathbf{x} * (\mathbf{y} * \mathbf{z})$$

qui est équivalent à dire que

$$L(\mathbf{x} * \mathbf{y})\mathbf{z} = L(\mathbf{x})(L(\mathbf{y})\mathbf{z}) = L(\mathbf{x})L(\mathbf{y})\mathbf{z}.$$

Comme cela est vrai pour tout $\mathbf{z}$, cela revient encore à dire que

$$L(\mathbf{x} * \mathbf{y}) = L(\mathbf{x})L(\mathbf{y}).$$

Définition 3.1.4 *Soit une algèbre $(V, *)$ et un sous-espace $U \subseteq V$. $(U, *)$ est une sous-algèbre si U est fermé pour $*$. Plus généralement, U est une sous-algèbre de V si U est isomorphe à une sous-algèbre de V.*

Définition 3.1.5 *La sous-algèbre engendrée par $\mathbf{x}$, notée $V(\mathbf{x})$ est la plus petite sous-algèbre contenant $\mathbf{x}$. Plus généralement, si S est un sous-ensemble de V, alors $V(S)$ désigne la plus petite sous-algèbre qui contient S.*

Pour tout $\mathbf{x} \in V$, il est facile de caractériser la sous-algèbre $(V(\mathbf{x}), *)$ en notant que $V(\mathbf{x})$ doit contenir un nombre minimal d'éléments de sorte qu'il contienne $\mathbf{x}$, qu'il soit un sous-espace vectoriel de V et qu'il soit fermé pour l'opération $*$. Ainsi, $V(\mathbf{x})$ devra contenir tout d'abord, $\alpha\mathbf{x}\ \forall \alpha \in I\!R$. Ensuite, en multipliant (au sens de $*$) tous les éléments déjà présents dans $V(\mathbf{x})$, le résultat doit encore appartenir à $V(\mathbf{x})$. Enfin, formant toutes les combinaisons linéaires possibles, le résultat doit appartenir à $V(\mathbf{x})$. En continuant ce procédé de multiplier et de former des combinaisons linéaires jusqu'à ce qu'il n'y ait plus de nouvel élément créé, nous parvenons à reconstituer l'ensemble $V(\mathbf{x})$.

Définition 3.1.6 *Soit une algèbre $(V, *)$. Si, $\forall \mathbf{x} \in V$, l'algèbre $V(\mathbf{x})$ est associative, alors $(V, *)$ est qualifiée d'algèbre associative par puissance.*

Pour montrer qu'une algèbre est associative par puissance il suffit de montrer que dans le produit $\mathbf{x} * \mathbf{x} * \ldots * \mathbf{x}$, l'ordre dans lequel les multiplications sont effectuées n'a pas d'importance ; dans ce cas, l'élément obtenu par k compositions

de $\mathbf{x}$ avec lui-même peut se noter sans ambiguïté $\mathbf{x}^k$, et $V(\mathbf{x})$ s'écrit

$$V(\mathbf{x}) = \left\{ \mathbf{v} \in V \,|\, \mathbf{v} = \sum_{i=0}^{k} \alpha_i \mathbf{x}^i,\ \alpha_i \in \mathbb{R},\ k \in \mathbb{N} \text{ ou } k = \infty \right\}.$$

3.1.2 Algèbres de Jordan

La notion d'algèbre de Jordan fait référence au physicien allemand Pascal Jordan (1902-1980) qui fut célèbre pour ses travaux avec Max Born en mécanique quantique, et non au mathématicien français du 19ème siècle Camille Jordan qui est à l'origine des fameux "blocs de Jordan" et du "théorème de la courbe fermée de Jordan".

Définition 3.1.7 *Soit $(\mathcal{J}, \circ)$ une algèbre. $(\mathcal{J}, \circ)$ est une algèbre de Jordan si*

$$\mathbf{x} \circ \mathbf{y} = \mathbf{y} \circ \mathbf{x}, \quad \forall \mathbf{x}, \mathbf{y} \in \mathcal{J} \tag{3.1}$$

$$\mathbf{x} \circ (\mathbf{x}^2 \circ \mathbf{y}) = \mathbf{x}^2 \circ (\mathbf{x} \circ \mathbf{y}), \quad \forall \mathbf{x}, \mathbf{y} \in \mathcal{J} \tag{3.2}$$

où $\mathbf{x}^2 = \mathbf{x} \circ \mathbf{x}$.

Donc, une algèbre de Jordan est une algèbre commutative qui possède une propriété similaire (mais plus faible) à l'associativité. Remarquons qu'en utilisant la transformation $L(\cdot)$, la propriété (3.2) revient à dire que

$$L(\mathbf{x})L(\mathbf{x}^2)\mathbf{y} = L(\mathbf{x}^2)L(\mathbf{x})\mathbf{y},\ \forall \mathbf{x}, \mathbf{y} \in \mathcal{J}$$

ou encore, puisque cela est valable pour tout $\mathbf{y} \in \mathcal{J}$,

$$L(\mathbf{x})L(\mathbf{x}^2) = L(\mathbf{x}^2)L(\mathbf{x}),\ \forall \mathbf{x} \in \mathcal{J}.$$

Par conséquent, une algèbre caractérisée par une matrice $L(\cdot)$, est une algèbre de Jordan si elle est commutative et si les matrices $L(\mathbf{x})$ et $L(\mathbf{x}^2)$ commutent pour tout $\mathbf{x} \in \mathcal{J}$.

Reprenons les trois exemples de la section précédente et, pour chacun d'eux, voyons s'ils correspondent ou non à une algèbre de Jordan.

1) Il est clair que l'algèbre $(\mathcal{M}_n, .)$ où . dénote la multiplication matricielle classique n'est pas une algèbre de Jordan puisque la multiplication matricielle n'est pas commutative.

2) $(\mathcal{M}_n, \circ)$ est une algèbre de Jordan. En effet, par définition de $\circ$ et par le fait que l'addition des matrices est commutative, il est clair que la propriété (3.1) est vérifiée. Ensuite, montrons que la propriété (3.2) est satisfaite. Notons

avant toute chose que pour tout $X \in \mathcal{M}_n$, $X^{\circ 2} := X \circ X = \frac{XX+XX}{2} = X^2$ (carré de la matrice X au sens classique). Ainsi, nous avons $\forall X \in \mathcal{M}_n$:

$$\begin{aligned} X \circ (X^{\circ 2} \circ Y) &= \frac{X(\frac{X^2Y+YX^2}{2})+(\frac{X^2Y+YX^2}{2})X}{2} \\ &= \frac{(\frac{X^3Y+XYX^2+X^2YX+YX^3}{2})}{2} \\ &= \frac{(\frac{X^3Y+X^2YX+XYX^2+YX^3}{2})}{2} \\ &= \frac{X^2(\frac{XY+YX}{2})+(\frac{XY+YX}{2})X^2}{2} \\ &= X^{\circ 2} \circ (X \circ Y). \end{aligned}$$

3) L'algèbre des formes quadratiques $(\mathbb{R}^{n+1}, \circ)$ est également une algèbre de Jordan. Ici aussi, il est clair en vertu de la définition de $\circ$, que (3.1) est vérifiée. Pour prouver que l'égalité (3.2) est satisfaite, nous allons montrer que $L(\mathbf{x})$ et $L(\mathbf{x}^2)$ commutent $\forall \mathbf{x} \in \mathbb{R}^{n+1}$. Notons tout d'abord, que d'après la définition de $\circ$, $\mathbf{x}^2$ vaut

$$\begin{pmatrix} x_0^2 + \bar{\mathbf{x}}^T B \bar{\mathbf{x}} \\ 2x_0 \bar{\mathbf{x}} \end{pmatrix}.$$

Ainsi, $\forall \mathbf{x} \in \mathbb{R}^{n+1}$, $L(\mathbf{x}^2)$ vaut

$$\begin{pmatrix} x_0^2 + \bar{\mathbf{x}}^T B \bar{\mathbf{x}} & 2x_0 \bar{\mathbf{x}}^T B \\ 2x_0 \bar{\mathbf{x}} & x_0^2 I + \bar{\mathbf{x}}^T B \bar{\mathbf{x}} I \end{pmatrix}$$

et

$$L(\mathbf{x})L(\mathbf{x}^2) = \begin{pmatrix} x_0^3 + 3x_0 \bar{\mathbf{x}}^T B \bar{\mathbf{x}} & (3x_0^2 + \bar{\mathbf{x}}^T B \bar{\mathbf{x}}) \bar{\mathbf{x}}^T B \\ (3x_0^2 + \bar{\mathbf{x}}^T B \bar{\mathbf{x}}) \bar{\mathbf{x}} & 2x_0 \bar{\mathbf{x}} \bar{\mathbf{x}}^T B + (x_0^3 + x_0 \bar{\mathbf{x}}^T B \bar{\mathbf{x}}) I \end{pmatrix} = L(\mathbf{x}^2)L(\mathbf{x}).$$

Donc $(\mathbb{R}^{n+1}, \circ)$ est bien une algèbre de Jordan.

3.2 Algèbre SOCP

3.2.1 Un cadre particulier

Pour définir l'algèbre du cône du second ordre, il nous faut préciser l'espace vectoriel V ainsi que la loi de composition entre deux éléments quelconques de V. Pour cela, nous allons nous baser sur l'algèbre des formes quadratiques $(\mathbb{R}^n, \circ)$ qui est une algèbre de Jordan non-associative et caractérisée par la matrice

$$L(\mathbf{x}) = \begin{pmatrix} x_0 & \bar{\mathbf{x}}^T B \\ \bar{\mathbf{x}} & x_0 I \end{pmatrix} \in \mathbb{R}^{n \times n},$$

où B est une matrice symétrique $(n-1) \times (n-1)$.

Définition 3.2.1 *L'algèbre du cône du second ordre de dimension n (Algèbre SOCP) correspond à l'algèbre des formes quadratiques $(\mathbb{R}^n, \circ)$ où la matrice B vaut I.*

Par cette définition, il est clair que l'algèbre du cône du second ordre est une algèbre de Jordan. Etant donné que cette algèbre a été particularisée avec $B \equiv I$, ses propriétés seront également plus particulières. Une propriété immédiate qui découle de cette particularisation est que la matrice $L(\cdot) \in I\!R^{n\times n}$ devient symétrique et s'écrit

$$L(\mathbf{x}) = \begin{pmatrix} x_0 & \bar{\mathbf{x}}^T \\ \bar{\mathbf{x}} & x_0 I \end{pmatrix} \quad \forall \mathbf{x} = (x_0; \bar{\mathbf{x}}) \in I\!R^n.$$

Nous nous apercevons que cette matrice coïncide avec la matrice $Arw(\mathbf{x})$ qui a été définie dans la section 1.2 et qui, comme nous l'avons vu, permet de caractériser l'ensemble des vecteurs qui appartiennent au cône du second ordre. Par conséquent, le fait d'avoir choisi $B \equiv I$ pour définir l'algèbre du cône du second ordre est tout-à-fait justifié. En outre, $\forall \mathbf{x}, \mathbf{y} \in I\!R^n$, nous aurons

$$\mathbf{x} \circ \mathbf{y} = Arw(\mathbf{x})\mathbf{y} = \begin{pmatrix} \mathbf{x}^T\mathbf{y} \\ x_0\bar{\mathbf{y}} + y_0\bar{\mathbf{x}} \end{pmatrix} = Arw(\mathbf{x})Arw(\mathbf{y})\mathbf{e}$$

avec $\mathbf{e} = (\mathbf{1}; \mathbf{0})$.

Qu'en est-il de l'associativité de $\circ$? Dans notre cadre particulier que constitue l'algèbre SOCP, nous allons montrer que l'associativité est vérifiée pour $n \leq 2$ mais plus pour $n > 2$.
Pour cela, rappelons que démontrer l'associativité revient à démontrer l'égalité entre $L(\mathbf{x} \circ \mathbf{y})$ et $L(\mathbf{x})L(\mathbf{y})$ pour tout $\mathbf{x}, \mathbf{y} \in I\!R^n$. Avec $\mathbf{x}, \mathbf{y} \in I\!R^n$, les matrices $L(\mathbf{x} \circ \mathbf{y})$ et $L(\mathbf{x})L(\mathbf{y})$ s'écrivent respectivement

$$\begin{pmatrix} \mathbf{x}^T\mathbf{y} & x_0\bar{\mathbf{y}}^T + y_0\bar{\mathbf{x}}^T \\ x_0\bar{\mathbf{y}} + y_0\bar{\mathbf{x}} & (\mathbf{x}^T\mathbf{y})I \end{pmatrix} \text{ et } \begin{pmatrix} \mathbf{x}^T\mathbf{y} & x_0\bar{\mathbf{y}}^T + y_0\bar{\mathbf{x}}^T \\ x_0\bar{\mathbf{y}} + y_0\bar{\mathbf{x}} & \bar{\mathbf{x}}\bar{\mathbf{y}}^T + (x_0y_0)I \end{pmatrix}.$$

D'où, l'égalité entre ces deux matrices se resume à l'égalité entre les sous-matrices :

$$(\mathbf{x}^T\mathbf{y})I \quad \text{et} \quad \bar{\mathbf{x}}\bar{\mathbf{y}}^T + (x_0y_0)I$$

ou encore, entre

$$(\bar{\mathbf{x}}^T\bar{\mathbf{y}})I \quad \text{et} \quad \bar{\mathbf{x}}\bar{\mathbf{y}}^T.$$

Or, cette égalité est toujours vraie pour $n = 1$ ou 2, mais plus (en général) pour $n > 2$.

Cependant, nous verrons dans la section suivante que l'associativité par puissance est satisfaite dans tous les cas.

3.2.2 Décomposition spectrale

L'algèbre linéaire nous garantit qu'une matrice symétrique $A \in \mathbb{R}^{n\times n}$ peut s'écrire sous la forme

$$A = Q\Lambda Q^T = \sum_{i=1}^{n} \lambda_i q_i q_i^T$$

où $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n) \in \mathbb{R}^{n\times n}$ et $Q = (q_1, \ldots, q_n) \in \mathbb{R}^{n\times n}$. λ_i est la i-ème valeur propre de A et q_i le vecteur propre correspondant de sorte que $QQ^T = Q^TQ = I$. Puisque A est symétrique, toutes les valeurs propres λ_i sont réelles. Supposons qu'il y ait k valeurs distinctes parmi les n valeurs propres, disons $\lambda^1, \ldots, \lambda^k$, que nous pouvons ordonner de la façon suivante :

$$\lambda^1 > \lambda^2 > \ldots > \lambda^k.$$

Avec n_i désignant la multiplicité de λ^i, définissons

$$P_i = \sum_{j=1}^{n_i} q_{i_j} q_{i_j}^T, \quad i = 1, \ldots, k,$$

tel que A s'écrive comme

$$A = \sum_{i=1}^{k} \lambda^i P_i.$$

La somme qui constitue le second membre de cette égalité est appelée la *décomposition spectrale* de A. Enonçons trois propriétés concernant les P_i qui entrent en jeu dans cette décomposition.

- $\forall i, j = 1, \ldots, k,\ i \neq j,\ P_iP_j = 0.$
  En effet, puisque $Q^TQ = I$, nous avons $\forall i, j = 1, \ldots, n,\ q_i^Tq_j = \delta_{ij}$ qui est le symbole de Kroenecker. D'où, si $i \neq j$

$$P_iP_j = \sum_{l=1}^{n_i} q_{i_l} q_{i_l}^T \sum_{k=1}^{n_j} q_{j_k} q_{j_k}^T = \sum_{l=1}^{n_i}\sum_{k=1}^{n_j} q_{i_l} \underbrace{(q_{i_l}^T q_{j_k})}_{=\delta_{i_l,j_k}=0} q_{j_k}^T = 0.$$

- $\sum_{i=1}^{k} P_i = I$. En effet, en notant que $n = \sum_{i=1}^{k} n_i$, nous avons

$$\sum_{i=1}^{k} P_i = \sum_{i=1}^{k}\sum_{j=1}^{n_i} q_{i_j} q_{i_j}^T = \sum_{l=1}^{n} q_l q_l^T = QQ^T = I.$$

- $\forall i = 1, \ldots, k,\ P_i^2 = P_i$. En effet,

$$P_i^2 = \sum_{j=1}^{n_i} q_{i_j} q_{i_j}^T \sum_{k=1}^{n_i} q_{i_k} q_{i_k}^T = \sum_{j=1}^{n_i}\sum_{k=1}^{n_i} q_{i_j} \underbrace{(q_{i_j}^T q_{i_k})}_{=\delta_{jk}} q_{i_k}^T = \sum_{j=1}^{n_i} q_{i_j} q_{i_j}^T = P_i.$$

Une telle décomposition spectrale est possible pour l'algèbre SOCP et les propriétés qui en découlent seront tout-à-fait analogues à celles des matrices P_i pour l'algèbre des matrices symétriques.

Définition 3.2.2 *Le polynôme*

$$p(\lambda, \mathbf{x}) = \lambda^2 - 2x_0\lambda + (x_0^2 - \|\bar{\mathbf{x}}\|^2)$$

est appelé polynôme caractéristique de $\mathbf{x}$ *et ses racines* $x_0 \pm \|\bar{\mathbf{x}}\|$ *sont appelées les valeurs propres de* $\mathbf{x}$.

Les polynômes caractéristiques et les valeurs propres dans cette algèbre joueront le même rôle que leurs homologues en algèbre des matrices symétriques si ce n'est que la situation est plus simple ici ; chaque élément de $I\!R^n$ a seulement deux valeurs propres réelles et celles-ci peuvent êtres calculées aisément.
Examinons à présent l'identité suivante pour un vecteur différent d'un multiple de l'identité e :

$$\mathbf{x} = \frac{1}{2}(x_0 + \|\bar{\mathbf{x}}\|)\begin{pmatrix} 1 \\ \frac{\bar{\mathbf{x}}}{\|\bar{\mathbf{x}}\|} \end{pmatrix} + \frac{1}{2}(x_0 - \|\bar{\mathbf{x}}\|)\begin{pmatrix} 1 \\ -\frac{\bar{\mathbf{x}}}{\|\bar{\mathbf{x}}\|} \end{pmatrix} \tag{3.3}$$

et pour un multiple de l'identité, i.e., $\mathbf{x} = (x_0; \mathbf{0})$,

$$\mathbf{x} = \frac{1}{2}x_0\begin{pmatrix} 1 \\ \frac{\bar{\mathbf{y}}}{\|\bar{\mathbf{y}}\|} \end{pmatrix} + \frac{1}{2}x_0\begin{pmatrix} 1 \\ -\frac{\bar{\mathbf{y}}}{\|\bar{\mathbf{y}}\|} \end{pmatrix}$$

pour un $\mathbf{y} = (y_0; \bar{\mathbf{y}}) \in I\!R^n$ différent d'un multiple de l'identité.

Définissons

$$\mathbf{c}_1 = \frac{1}{2}\begin{pmatrix} 1 \\ \frac{\bar{\mathbf{x}}}{\|\bar{\mathbf{x}}\|} \end{pmatrix};\ \mathbf{c}_2 = \frac{1}{2}\begin{pmatrix} 1 \\ -\frac{\bar{\mathbf{x}}}{\|\bar{\mathbf{x}}\|} \end{pmatrix};\ \lambda_1 = x_0 + \|\bar{\mathbf{x}}\|;\ \lambda_2 = x_0 - \|\bar{\mathbf{x}}\|. \tag{3.4}$$

Il suit que (3.3) peut s'écrire comme

$$\mathbf{x} = \lambda_1\mathbf{c}_1 + \lambda_2\mathbf{c}_2. \tag{3.5}$$

Cette égalité constitue la décomposition spectrale de $\mathbf{x}$.
Enonçons à présent les premières propriétés liées à cette décomposition qui témoignerons de l'analogie qu'il y a ici avec l'algèbre des matrices symétriques.

$$\mathbf{c}_1 \circ \mathbf{c}_2 = \mathbf{0} \tag{3.6}$$

$$\mathbf{c}_1 + \mathbf{c}_2 = \mathbf{e} \tag{3.7}$$

$$\mathbf{c}_1^2 = \mathbf{c}_1 \text{ et } \mathbf{c}_2^2 = \mathbf{c}_2 \tag{3.8}$$

$$\mathbf{c}_1 = R\mathbf{c}_2 \text{ et } \mathbf{c}_2 = R\mathbf{c}_1 \tag{3.9}$$

$$\mathbf{c}_1, \mathbf{c}_2 \in \text{bd}\,\mathcal{Q} \tag{3.10}$$

En effet,

- $\mathbf{c}_1 \circ \mathbf{c}_2 = \begin{pmatrix} \mathbf{c}_1{}^T\mathbf{c}_2 \\ \frac{1}{2}\bar{\mathbf{c}}_2 + \frac{1}{2}\bar{\mathbf{c}}_1 \end{pmatrix} = \begin{pmatrix} 0 \\ \mathbf{0} \end{pmatrix} \Rightarrow$ (3.6).
- (3.7) est evident d'après la définition.
- Pour $i = 1$, par exemple, nous avons par (3.6) et (3.7),

$$\mathbf{c}_1 = \mathbf{c}_1 \circ (\mathbf{c}_1 + \mathbf{c}_2) = \mathbf{c}_1 \circ \mathbf{c}_1 = \mathbf{c}_1^2 \Rightarrow (3.8).$$

- Pour obtenir (3.9), souvenons-nous que la matrice R définie à la section 1.1, équivaut à $\begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & -I \end{pmatrix}$. Cela nous permet d'écrire

$$R\mathbf{c}_2 = \begin{pmatrix} \frac{1}{2} \\ -\frac{1}{2} I\left(\frac{-\bar{\mathbf{x}}}{\|\bar{\mathbf{x}}\|}\right) \end{pmatrix} = \mathbf{c}_1.$$

  L'autre égalité s'établit en pré-multipliant l'égalité obtenue par R.
- (3.10) s'obtient en remarquant que $\mathbf{c}_1$ et $\mathbf{c}_2$ vérifient une inégalité du type $\mathbf{x} = (x_0; \bar{\mathbf{x}}) \succcurlyeq_{\mathcal{Q}} 0$ avec une égalité et appartiennent donc à bd $\mathcal{Q}$.

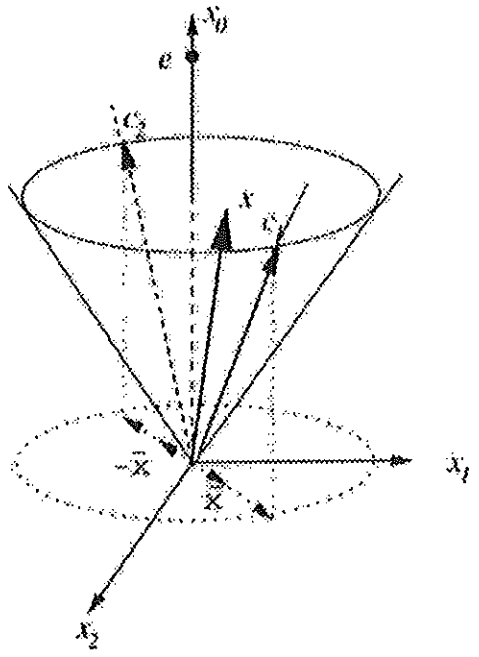


figure 3.1: Structure de Jordan d'un vecteur $\mathbf{x}$ appartenant au cône du second ordre

Toute paire de vecteurs $\{\mathbf{c}_1, \mathbf{c}_2\}$ qui satisfait les propriétés (3.6), (3.7) et (3.8) est appelée une *structure de Jordan.* Un vecteur $\mathbf{c}_i$ de cette paire joue un rôle

tout-à-fait analogue à celui d'une matrice P_i qui intervient dans la décomposition spectrale d'une matrice symétrique comme nous l'avons vu précédemment. Nous sommes à présent en mesure de prouver que l'algèbre $(\mathbb{R}^n, \circ)$ est associative par puissance.

Proposition 3.2.1 *L'algèbre SOCP $(\mathbb{R}^n, \circ)$ est associative par puissance.*

Preuve : Soit $\mathbf{x} = \lambda_1 \mathbf{c}_1 + \lambda_2 \mathbf{c}_2$ la décomposition spectrale de $\mathbf{x}$. Pour $p \in \mathbb{N}$, définissons

$$\mathbf{x}^{\circ 0} = \mathbf{e}, \quad \mathbf{x}^{\circ p} = \mathbf{x} \circ \mathbf{x}^{\circ(p-1)}, \quad \mathbf{x}^p = \lambda_1^p \mathbf{c}_1 + \lambda_2^p \mathbf{c}_2.$$

Tout d'abord, montrons par récurrence que $\forall p \in \mathbb{N} : \mathbf{x}^{\circ p} = \mathbf{x}^p$.
L'égalité est vérifiée pour $p = 0$ en vertu de la relation (3.7). Supposons que l'égalité soit vraie pour $k \in \mathbb{N}$. Nous avons alors

$$\begin{aligned} \mathbf{x}^{\circ(k+1)} &= \mathbf{x} \circ \mathbf{x}^{\circ k} = (\lambda_1 \mathbf{c}_1 + \lambda_2 \mathbf{c}_2) \circ (\lambda_1^k \mathbf{c}_1 + \lambda_2^k \mathbf{c}_2) \\ &= \lambda_1^{k+1} \mathbf{c}_1^2 + \lambda_2^{k+1} \mathbf{c}_2^2 + \lambda_1 \lambda_2^k \mathbf{c}_1 \circ \mathbf{c}_2 + \lambda_1^k \lambda_2 \mathbf{c}_2 \circ \mathbf{c}_1 \\ &= \lambda_1^{k+1} \mathbf{c}_1 + \lambda_2^{k+1} \mathbf{c}_2 \\ &= \mathbf{x}^{k+1} \end{aligned}$$

où l'avant-dernière égalité est obtenue grâce à (3.6) et (3.8).

Ensuite, pour des entiers q, r, s et des réels α_i, β_j, γ_k $(i = 0, \ldots, q,\ j = 0, \ldots, r,\ k = 0, \ldots, s)$ tous arbitraires, posons

$$\mathbf{u} = \sum_{i=0}^{q} \alpha_i \mathbf{x}^i, \quad \mathbf{v} = \sum_{j=0}^{r} \beta_j \mathbf{x}^j, \quad \mathbf{w} = \sum_{k=0}^{s} \gamma_k \mathbf{x}^k.$$

$$\begin{aligned} \mathbf{u} \circ \mathbf{v} &= \left(\textstyle\sum_{i=0}^{q} \alpha_i \mathbf{x}^i\right) \circ \left(\textstyle\sum_{j=0}^{r} \beta_j \mathbf{x}^j\right) \\ &= \left(\textstyle\sum_{i=0}^{q} \alpha_i (\lambda^i \mathbf{c}_1 + \lambda^i \mathbf{c}_2)\right) \circ \left(\textstyle\sum_{j=0}^{r} \beta_j (\lambda^j \mathbf{c}_1 + \lambda^j \mathbf{c}_2)\right) \\ &= \textstyle\sum_{i=0}^{q} \sum_{j=0}^{r} \alpha_i \beta_j \lambda^{i+j} \mathbf{c}_1 + \sum_{i=0}^{q} \sum_{j=0}^{r} \alpha_i \beta_j \lambda^{i+j} \mathbf{c}_2 \end{aligned}$$

Ainsi, après avoir établi cette expression pour $\mathbf{u} \circ \mathbf{v}$, il est clair que

$$(\mathbf{u} \circ \mathbf{v}) \circ \mathbf{w} = \sum_{i,j,k} \alpha_i \beta_j \gamma_k \lambda^{i+j+k} \mathbf{c}_1 + \sum_{i,j,k} \alpha_i \beta_j \gamma_k \lambda^{i+j+k} \mathbf{c}_2 = \mathbf{u} \circ (\mathbf{v} \circ \mathbf{w})$$

et donc, en accord avec la définition 3.1.6, nous avons établi que $(\mathbb{R}^n, \circ)$ est associative par puissance.

□

Cette proposition va nous permettre de définir les notions de degré et de rang qui seront utilisées pour justifier l'appellation cône "du second ordre".

3.2.3 Notions de degré et rang.

Définition 3.2.3 *Soit $(V, *)$ une algèbre associative par puissance dont l'élément identité est* e *et telle que $dim(V) = n$. Pour chaque élément* $\mathbf{x} \in (V, *)$, *le degré de* x *est défini comme étant le plus petit entier d tel que l'ensemble* $\{\mathbf{e}, \mathbf{x}, \ldots, \mathbf{x}^d\}$ *soit linéairement dépendant dans* $(V, *)$.

Puisqu'un ensemble de plus de n vecteurs dans un espace vectoriel de dimension n est toujours linéairement dépendant, il est clair que : $d \leq n$.

Définition 3.2.4 *Soit* $r = \max_{\mathbf{x} \in V} deg(\mathbf{x})$. *Alors, r est appelé le rang de V (sous-entendu $(V, *)$) et noté $rang(V)$.*

Définition 3.2.5 *Tout élément de V est dit régulier si* $deg(\mathbf{x}) = rang(V)$, *et non-régulier, sinon.*

Dans l'algèbre SOCP, tout élément x vérifie l'identité vectorielle suivante :

$$\mathbf{x}^2 - 2x_0\mathbf{x} + (x_0^2 - \|\bar{\mathbf{x}}\|^2)\mathbf{e} = \mathbf{0} \tag{3.11}$$

qui s'obtient directement en notant que $\mathbf{x}^2 = (\|\mathbf{x}\|^2; 2x_0\bar{\mathbf{x}})$. Cela nous permet d'affirmer que le rang de l'algèbre SOCP $(\mathbb{R}^n, \circ)$ ne peut excéder 2. En outre, si $\mathbf{x}^*$ est un élément de $\mathbb{R}^n$ différent d'un multiple de l'identité e (donc forcément non nul) alors, par définition, $deg(\mathbf{x}^*) \geq 2$ et donc, en vertu de (3.11), $deg(\mathbf{x}^*) = 2$. Par conséquent, l'algèbre SOCP est de rang 2, c'est pourquoi le cône qui est basé sur cette algèbre particulière est dit du second ordre.

Notons que les vecteurs de l'algèbre SOCP $(\mathbb{R}^n, \circ)$ qualifiés de non-réguliers peuvent être aisément identifiés. En effet, ces vecteurs doivent forcément être de degré 1 ou 0 et donc linéairement dépendants avec e. D'où, les éléments non-réguliers de $(\mathbb{R}^n, \circ)$ correspondent aux multiples de e.

Définition 3.2.6 *Soit $p(\lambda, \mathbf{x})$ le polynôme caractéristique d'un vecteur* $\mathbf{x} \in \mathbb{R}^n$. *La trace et le déterminant de* x, *notés $tr(\mathbf{x})$ et* $\det(\mathbf{x})$ *sont définis comme étant respectivement l'opposé du coefficient de λ et le terme indépendant dans $p(\lambda, \mathbf{x})$, i.e.,*

$$tr(\mathbf{x}) = 2x_0 \quad et \quad \det(\mathbf{x}) = x_0^2 - \|\bar{\mathbf{x}}\|^2.$$

Remarquons immédiatement que, similairement aux matrices, nous avons

$$tr(\mathbf{x}) = \lambda_1 + \lambda_2 = 2x_0 \quad \text{et} \quad \det(\mathbf{x}) = \lambda_1\lambda_2 = x_0^2 - \|\bar{\mathbf{x}}\|^2.$$

Une autre façon de caractériser les vecteurs appartenant au cône du second ordre (ou à son intérieur) est donnée par les équivalences suivantes que l'on établit immédiatement d'après la définition des valeurs propres d'un vecteur $\mathbf{x}$

$$\mathbf{x} \in \mathcal{Q} \Leftrightarrow \lambda_1, \lambda_2 \geq 0, \quad \mathbf{x} \in \operatorname{int} \mathcal{Q} \Leftrightarrow \lambda_1, \lambda_2 > 0.$$

Aussi, par analogie aux matrices symétriques, tout vecteur $\mathbf{x} \in \mathcal{Q}$ sera qualifié de *semi-défini positif* et tout vecteur $\mathbf{x} \in \operatorname{int} \mathcal{Q}$ sera qualifié de *défini positif*. De même, tout vecteur $\mathbf{x}$ tel que $det(\mathbf{x}) \neq 0 (= 0)$ sera dit *non-singulier* (*singulier*). Soulignons que les éléments réguliers ne doivent pas être confondus avec les éléments non-singuliers, puisque un élément régulier peut être aussi bien non-singulier (ex : $(1,2) \in I\!R^2$) que singulier (ex : $(1,1) \in I\!R^2$) ; de même, un élément non-singulier peut être aussi bien régulier que non-régulier (ex : l'identité $\mathbf{e}$).

La décomposition spectrale d'un élément de $(I\!R^n, \circ)$ a l'avantage de permettre l'extension de toute fonction continue à valeur réelle à l'algèbre SOCP. Elle permet également de définir aisément les puissances d'éléments de $(I\!R^n, \circ)$ ainsi que des normes analogues aux normes matricielles, comme par exemple, la norme de Frobénius et la norme 2 (induite par le produit scalaire).

Définition 3.2.7 *Soit* $\mathbf{x} \in I\!R^n$ *et soit* $\mathbf{x} = \lambda_1 \mathbf{c}_1 + \lambda_2 \mathbf{c}_2$ *sa décomposition spectrale. Alors, nous définissons pour* $\mathbf{x}$ *:*

$$\|\mathbf{x}\|_F = \sqrt{\lambda_1^2 + \lambda_2^2}, \tag{3.12}$$

$$\|\mathbf{x}\|_2 = \max\{|\lambda_1|, |\lambda_2|\}, \tag{3.13}$$

$$\mathbf{x}^{-1} = \lambda_1^{-1}\mathbf{c}_1 + \lambda_2^{-1}\mathbf{c}_2 \quad si\ \det(\mathbf{x}) \neq 0, \tag{3.14}$$

$$\mathbf{x}^{1/2} = \lambda_1^{1/2}\mathbf{c}_1 + \lambda_2^{1/2}\mathbf{c}_2 \ pour\ \mathbf{x} \succcurlyeq_{\mathcal{Q}} 0, \tag{3.15}$$

$$f(\mathbf{x}) = f(\lambda_1)\mathbf{c}_1 + f(\lambda_2)\mathbf{c}_2 \quad si\ f(\lambda_i)\ est\ bien\ défini\ pour\ i = 1,2, \tag{3.16}$$

$$\mathbf{x}^{-p} = (\mathbf{x}^{-1})^p = (\mathbf{x}^p)^{-1} = \lambda_1^{-p}\mathbf{c}_1 + \lambda_2^{-p}\mathbf{c}_2, \tag{3.17}$$

si $\lambda_1^{-p}, \lambda_2^{-p}$ *sont bien définis.*

La norme de Frobénius définie en (3.12) peut être réécrite en fonction de la norme euclidienne du vecteur $\mathbf{x}$. En effet

$$\|\mathbf{x}\|_F = \sqrt{\lambda_1^2 + \lambda_2^2} = \sqrt{(x_0 + \|\bar{\mathbf{x}}\|)^2 + (x_0 - \|\bar{\mathbf{x}}\|)^2} = \sqrt{2(x_0^2 + \|\bar{\mathbf{x}}\|^2)} = \sqrt{2}\|\mathbf{x}\|.$$

De plus, pour tout $\mathbf{x}$ nonsingulier, $\mathbf{x}^{-1}$ est appelé *inverse* de $\mathbf{x}$. Cet inverse vérifie les relations suivantes

$$\mathbf{x}^{-1} \circ \mathbf{x} = \mathbf{e} \quad \text{et} \quad \mathbf{x}^{-1} = R\mathbf{x}/\det(\mathbf{x}).$$

En effet, grâce à la décomposition spectrale de $\mathbf{x}$ nous avons

$$\begin{aligned}
\mathbf{x}^{-1} \circ \mathbf{x} &= (\lambda_1^{-1}\mathbf{c}_1 + \lambda_2^{-1}\mathbf{c}_2) \circ (\lambda_1\mathbf{c}_1 + \lambda_2\mathbf{c}_2) \\
&= \lambda_1^{-1}\lambda_1\mathbf{c}_1 + \lambda_2^{-1}\lambda_2\mathbf{c}_2 \\
&= \mathbf{e}. \\
\mathbf{x}^{-1} &= \lambda_1^{-1}\mathbf{c}_1 + \lambda_2^{-1}\mathbf{c}_2 \\
&= \tfrac{\lambda_2}{\lambda_1\lambda_2} R\mathbf{c}_2 + \tfrac{\lambda_1}{\lambda_1\lambda_2} R\mathbf{c}_1 \\
&= \tfrac{1}{\lambda_1\lambda_2} R(\lambda_1\mathbf{c}_1 + \lambda_2\mathbf{c}_2) \\
&= R\mathbf{x}/\det(\mathbf{x}).
\end{aligned}$$

Proposition 3.2.2 *Soit $\mathbf{x} \in I\!R^n$ tel que $\mathbf{x}$ se décompose en $\mathbf{x} = \lambda_1\mathbf{c}_1 + \lambda_2\mathbf{c}_2$. Alors, $\forall p \in I\!N$, $\forall \alpha \in I\!R$ et $\forall \mathbf{y} \in I\!R^n$, nous avons les propriétés suivantes :*

1. $tr(\alpha\mathbf{x} + \beta\mathbf{y}) = \alpha tr(\mathbf{x}) + \beta tr(\mathbf{y})$,
2. $\det(\alpha\mathbf{x}) = \alpha^2 \det(\mathbf{x})$,
3. $\det(\mathbf{x}^p) = \det^p(\mathbf{x})$,
4. *Si $\mathbf{x}$ est non-singulier,* $\det(\mathbf{x}^{-1}) = \det^{-1}(\mathbf{x})$.

Preuve : 1) Le polynôme caractéristique de $\alpha\mathbf{x} + \beta\mathbf{y}$ s'écrit :

$$p(\lambda, \alpha\mathbf{x} + \beta\mathbf{y}) = \lambda^2 - 2(\alpha\mathbf{x} + \beta\mathbf{y})_0\lambda + (\alpha\mathbf{x} + \beta\mathbf{y})_0^2 - \|\alpha\bar{\mathbf{x}} + \beta\bar{\mathbf{y}}\|^2.$$

Nous avons alors, $tr(\alpha\mathbf{x}+\beta\mathbf{y}) = 2(\alpha\mathbf{x}+\beta\mathbf{y})_0 = \alpha 2x_0 + \beta 2y_0 = \alpha tr(\mathbf{x}) + \beta tr(\mathbf{y})$.

2) Le terme indépendant du polynôme caractéristique de $\alpha\mathbf{x}$ est

$$(\alpha\mathbf{x})_0^2 - \|\alpha\bar{\mathbf{x}}\|^2 = \alpha^2(x_0^2 - \|\bar{\mathbf{x}}\|^2),$$

d'où le résultat grâce à la définition du déterminant.

3) Nous avons vu dans la preuve de la proposition (3.2.1) que la puissance $\mathbf{x}^p$ est définie par $\mathbf{x}^p = \lambda_1^p\mathbf{c}_1 + \lambda_2^p\mathbf{c}_2$ où, pour $i = 1, 2$, $\mathbf{c}_i = \frac{1}{2}(1; \pm\mathbf{d})$ avec $\|\mathbf{d}\| = 1$. Cette décomposition spectrale nous permet d'écrire :

$$(\mathbf{x}^p)_0^2 = \frac{1}{4}(\lambda_1^p + \lambda_2^p)^2, \quad \|(\overline{\mathbf{x}^p})\|^2 = \|\frac{1}{2}(\lambda_1^p - \lambda_2^p)\mathbf{d}\|^2 = \frac{1}{4}(\lambda_1^p - \lambda_2^p)^2.$$

Ainsi, en vertu de la définition du déterminant, nous obtenons

$$\det(\mathbf{x}^p) = (\mathbf{x}^p)_0^2 - \|(\overline{\mathbf{x}^p})\|^2 = \lambda_1^p\lambda_2^p = \det^p(\mathbf{x}).$$

4) En nous basant sur le fait que $\mathbf{x}^{-1} = R\mathbf{x}/\det(\mathbf{x}) = (x_0; -\bar{\mathbf{x}})/\det(\mathbf{x})$, et en utilisant la propriété 2), nous obtenons

$$\det(\mathbf{x}^{-1}) = \frac{1}{\det^2(\mathbf{x})} \det((x_0; -\bar{\mathbf{x}})) = \frac{1}{\det^2(\mathbf{x})} \det(\mathbf{x}) = \det^{-1}(\mathbf{x}).$$

□

3.2.4 La représentation quadratique

En plus de $Arw(\mathbf{x})$, il existe une autre matrice très importante associée à tout $\mathbf{x}$, appelée *représentation quadratique.*

Définition 3.2.8 $\forall\, \mathbf{x} \in I\!R^n$, *la représentation quadratique, notée* $Q_{\mathbf{x}}$, *est une matrice* $n \times n$ *définie par*

$$Q_{\mathbf{x}} = 2Arw^2(\mathbf{x}) - Arw(\mathbf{x}^2). \tag{3.18}$$

Nous pouvons donner une expression plus explicite pour $Q_{\mathbf{x}}$; sachant que $\mathbf{x}^2 = (\|\mathbf{x}\|^2;\, 2x_0\bar{\mathbf{x}})$, nous obtenons :

$$Q_{\mathbf{x}} = \begin{pmatrix} \|\mathbf{x}\|^2 & 2x_0\bar{\mathbf{x}}^T \\ 2x_0\bar{\mathbf{x}} & \det(\mathbf{x})I + 2\bar{\mathbf{x}}\bar{\mathbf{x}}^T \end{pmatrix} = 2\mathbf{x}\mathbf{x}^T - \det(\mathbf{x})R,$$

ce qui nous donne pour tout $\mathbf{y} \in I\!R^n$,

$$Q_{\mathbf{x}}\mathbf{y} = 2(\mathbf{x}^T\mathbf{y})\mathbf{x} - \det(\mathbf{x})R\mathbf{y}$$

qui est un vecteur de fonctions quadratiques en $\mathbf{x}$. En fonction de l'opérateur $\circ$, nous obtenons à partir de (3.18),

$$\forall\, \mathbf{y} \in I\!R^n, \quad Q_{\mathbf{x}}\mathbf{y} = 2\mathbf{x} \circ (\mathbf{x} \circ \mathbf{y}) - \mathbf{x}^2 \circ \mathbf{y}.$$

La matrice de représentation semble n'être à première vue qu'un opérateur arbitraire mais, comme nous le verrons, elle joue un rôle extrêmement important dans les tous aspects de la programmation du cône du second ordre. En réalité, cette matrice peut-être perçue comme l'analogue de l'opérateur qui, à Y, fait correspondre XYX dans l'algèbre $(\mathcal{M}_n, \circ)$. En effet, $\forall X, Y \in (\mathcal{M}_n, \circ)$

$$\begin{aligned} Q_X Y &= 2X \circ (X \circ Y) - X^2 \circ Y \\ &= \left(X\tfrac{XY+YX}{2} + \tfrac{XY+YX}{2}X\right) - \tfrac{X^2Y+YX^2}{2} \\ &= \tfrac{X^2Y+XYX+XYX+YX^2-X^2Y-YX^2}{2} \\ &= XYX. \end{aligned}$$

Dans un chapitre ultérieur, nous ferons usage d'une matrice de représentation généralisée à deux éléments de $I\!R^n$, qui est notée $Q_{\mathbf{x},\mathbf{y}}$ et définie par

$$\forall\, \mathbf{x}, \mathbf{y} \in I\!R^n \quad Q_{\mathbf{x},\mathbf{y}} = Arw(\mathbf{x})Arw(\mathbf{y}) + Arw(\mathbf{y})Arw(\mathbf{x}) - Arw(\mathbf{x} \circ \mathbf{y})$$

avec $Q_{\mathbf{x},\mathbf{x}} = Q_{\mathbf{x}}$.

Les propriétés les plus fondamentales pour $Arw(\mathbf{x})$ et $Q_{\mathbf{x}}$ vont faire l'objet du théorème qui va suivre. Tout d'abord, considérons le lemme suivant issu de la théorie des matrices :

Lemme 3.2.1 *Soient X et Y deux matrices carrées $n \times n$. Alors, nous avons l'équivalence suivante :*

$$XY = YX \iff X \text{ et } Y \text{ ont les mêmes vecteurs propres.}$$

Théorème 3.2.1 *Soit $\mathbf{x} \in I\!R^n$ qui se décompose en $\mathbf{x} = \lambda_1 \mathbf{c}_1 + \lambda_2 \mathbf{c}_2$, avec $\lambda_1 \neq \lambda_2$. Les trois assertions suivantes sont vraies :*

1. *$Arw(\mathbf{x})$ et $Q_{\mathbf{x}}$ commutent et donc, possèdent les mêmes vecteurs propres.*
2. *$\lambda_1 = x_0 + \|\bar{\mathbf{x}}\|$ et $\lambda_2 = x_0 - \|\bar{\mathbf{x}}\|$ sont des valeurs propres de $Arw(\mathbf{x})$ de multiplicité 1 dont les vecteurs propres correspondant sont, respectivement, $\mathbf{c}_1$ et $\mathbf{c}_2$. De plus, x_0 est une valeur propre de $Arw(\mathbf{x})$ de multiplicité $n-2$.*
3. *$\lambda_1^2 = (x_0 + \|\bar{\mathbf{x}}\|)^2$ et $\lambda_2^2 = (x_0 - \|\bar{\mathbf{x}}\|)^2$ sont des valeurs propres de $Q_{\mathbf{x}}$ de multiplicité 1 dont les vecteurs propres correspondant sont, respectivement, $\mathbf{c}_1$ et $\mathbf{c}_2$. De plus, $det(\mathbf{x}) = x_0^2 - \|\bar{\mathbf{x}}\|^2 = \lambda_1 \lambda_2$ est une valeur propre de $Q_{\mathbf{x}}$ de multiplicité $n-2$.*

Preuve : 1) L'assertion découle du fait que l'algèbre SOCP est une algèbre de Jordan ; de ce fait, nous aurons pour tout $\mathbf{x} \in I\!R^n$

$$Arw(\mathbf{x})Arw(\mathbf{x}^2) = Arw(\mathbf{x}^2)Arw(\mathbf{x}).$$

Ainsi, puisque $Q_{\mathbf{x}}$ est une combinaison linéaire de $Arw(\mathbf{x})$ et de $Arw(\mathbf{x}^2)$, il est clair que $Arw(\mathbf{x})$ et $Q_{\mathbf{x}}$ commutent.

2) Pour cette assertion il nous suffit de montrer la propriété se rapportant aux vecteurs propres puisque l'étude des valeurs propres de $Arw(\mathbf{x})$ a déjà été effectuée dans la preuve de la proposition (1.2.1).
Nous allons montrer que $\mathbf{c}_1$ est un vecteur propre de $Arw(\mathbf{x})$ par rapport à λ_1 (le cas pour $\mathbf{c}_2$ est tout-à-fait similaire).

$$\begin{aligned} Arw(\mathbf{x})\mathbf{c}_1 &= \mathbf{x} \circ \mathbf{c}_1 \\ &= (\lambda_1 \mathbf{c}_1 + \lambda_2 \mathbf{c}_2) \circ \mathbf{c}_1 \\ &= \lambda_1 \mathbf{c}_1^2 + \lambda_2 (\mathbf{c}_2 \circ \mathbf{c}_1) \\ &= \lambda_1 \mathbf{c}_1. \end{aligned}$$

3) Nous calculons tout d'abord les valeurs propres de $Arw(\mathbf{x}^2)$.
En vertu du point 2), la valeur propre de multiplicité $n-2$ est donnée par $(\mathbf{x}^2)_0 = \|\mathbf{x}\|^2 = x_0^2 + \|\bar{\mathbf{x}}\|^2$. En outre, nous avons

$$\begin{aligned} Arw(\mathbf{x}^2)\mathbf{c}_1 = \mathbf{x}^2 \circ \mathbf{c}_1 = (\lambda_1^2 \mathbf{c}_1 + \lambda_2^2 \mathbf{c}_2) \circ \mathbf{c}_1 = \lambda_1^2 \mathbf{c}_1 \\ Arw(\mathbf{x}^2)\mathbf{c}_2 = \mathbf{x}^2 \circ \mathbf{c}_2 = (\lambda_1^2 \mathbf{c}_1 + \lambda_2^2 \mathbf{c}_2) \circ \mathbf{c}_2 = \lambda_2^2 \mathbf{c}_2. \end{aligned}$$

Par conséquent, λ_1^2 et λ_2^2 sont les valeurs propres de $Arw(\mathbf{x}^2)$ de multiplicité égale à 1 (sinon $Arw(\mathbf{x}^2)$ aurait plus de n valeurs propres) dont les vecteurs propres correspondant sont, respectivement, $\mathbf{c_1}$ et $\mathbf{c_2}$.
Considérons ensuite les factorisations de Jordan de $Arw(\mathbf{x})$ et $Arw(\mathbf{x}^2)$:

$$Arw(\mathbf{x}) = P\Lambda P^T \quad Arw(\mathbf{x}^2) = P\Omega P^T,$$

avec $P = (\sqrt{2}\mathbf{c_1}, \widehat{Q}, \sqrt{2}\mathbf{c_2})$ la matrice orthogonale des vecteurs propres de $Arw(\mathbf{x})$ et $Arw(\mathbf{x}^2)$, et

$$\Lambda = \begin{pmatrix} \lambda_1 & \mathbf{0}^{\mathrm{T}} & 0 \\ \mathbf{0} & x_0 I & \mathbf{0} \\ 0 & \mathbf{0}^T & \lambda_2 \end{pmatrix}, \quad \Omega = \begin{pmatrix} \lambda_1^2 & \mathbf{0}^{\mathrm{T}} & 0 \\ \mathbf{0} & (x_0^2 + \|\bar{\mathbf{x}}\|^2) I & \mathbf{0} \\ 0 & \mathbf{0}^T & \lambda_2^2 \end{pmatrix},$$

leur forme de Jordan.
Enfin, en utilisant sa définition, nous obtenons la factorisation suivante pour $Q_{\mathbf{x}}$

$$Q_{\mathbf{x}} = P\{2\Lambda^2 - \Omega\}P^T = P\Sigma P^T,$$

où la matrice des valeurs propre Σ est donnée par

$$\Sigma = \begin{pmatrix} \lambda_1^2 & \mathbf{0}^{\mathrm{T}} & 0 \\ \mathbf{0} & \det(\mathbf{x}) I & \mathbf{0} \\ 0 & \mathbf{0}^T & \lambda_2^2 \end{pmatrix}$$

ce qui termine la preuve de 3).

□

Remarques :

-Puisque le déterminant d'une matrice correspond au produit de ses valeurs propres, nous déduisons de 3) que la seule façon pour que $Q_{\mathbf{x}}$ soit non-singulière est que $\lambda_1.\lambda_2 \neq 0$.

-Lorsque la contrainte $\mathbf{x} \succcurlyeq_Q 0$ est présente, $\lambda_{\min}(Arw(\mathbf{x})) = \lambda_2 \geq 0$. Donc, dans ce cas, la non-singularité de $Arw(\mathbf{x})$ implique que $\lambda_2 > 0$ et donc, puisque $\lambda_1 \geq \lambda_2$, que $\mathbf{x}$ est non-singulier. Inversément, si $\mathbf{x}$ est non-singulier, λ_2 doit forcément être non nulle, et donc, strictement positive, ce qui garantit la non-singularité de $Arw(\mathbf{x})$.

Corollaire 3.2.1 *Soit* $\mathbf{x} \in I\!R^n$.

1. $Q_{\mathbf{x}}$ *est non-singulière si et seulement si* $\mathbf{x}$ *est non-singulier.*
2. *Si* $\mathbf{x} \succcurlyeq_Q 0$, $Arw(\mathbf{x})$ *est non-singulière si et seulement si* $\mathbf{x}$ *est non-singulier.*

3.2.5 Commutativité et structure de Jordan

Par définition, la loi $\circ$ est commutative. Cependant, il est important de d'introduire une notion de commutativité entre éléments considérés comme des opérateurs, et qui est analogue à la commutativité des matrices. Dans notre cadre que constitue l'algèbre SOCP, nous allons définir la commutativité en termes de structure de Jordan. Plus précisément, nous avons :

Définition 3.2.9 *Les éléments* $\mathbf{x}$ *et* $\mathbf{y}$ *considérés comme opérateurs, commutent si ils partagent une structure de Jordan commune, i.e.,*

$$\mathbf{x} = \lambda_1 \mathbf{c}_1 + \lambda_2 \mathbf{c}_2, \quad \mathbf{y} = \omega_1 \mathbf{c}_1 + \omega_2 \mathbf{c}_2$$

pour une structure de Jordan $\{\mathbf{c}_1, \mathbf{c}_2\}$.

Géométriquement, le fait que $\mathbf{x}$ et $\mathbf{y}$ commutent par rapport à une structure de Jordan $\{\mathbf{c}_1, \mathbf{c}_2\}$ revient à dire qu'ils appartiennent tous deux au sous-espace engendré par $\mathbf{c}_1$ et $\mathbf{c}_2$, noté $\mathrm{Span}\{\mathbf{c}_1, \mathbf{c}_2\}$. Intéressons-nous à présent aux manières de caractériser des vecteurs qui commutent.

Proposition 3.2.3 *Soit* $\mathbf{x} \in I\!R^n$.

Si $\bar{\mathbf{x}} = 0$, *alors* $\mathbf{x}$ *commute avec tout* $\mathbf{y} \in I\!R^n$.

Preuve : Soit $\mathbf{y} \in I\!R^n$ tel que $\mathbf{y}$ se décompose en $\mathbf{y} = \omega_1 \mathbf{c}_1 + \omega_2 \mathbf{c}_2$ avec $\mathbf{c}_1 = (\frac{1}{2}; \frac{1}{2}\mathbf{d})$ et $\mathbf{c}_2 = (\frac{1}{2}; -\frac{1}{2}\mathbf{d})$ tels que $\|\mathbf{d}\| = 1$. Le résultat est alors immédiat puisque

$$\mathbf{x} = x_0 \begin{pmatrix} \frac{1}{2} \\ \frac{1}{2}\mathbf{d} \end{pmatrix} + x_0 \begin{pmatrix} \frac{1}{2} \\ -\frac{1}{2}\mathbf{d} \end{pmatrix} = \lambda_1 \mathbf{c}_1 + \lambda_2 \mathbf{c}_2.$$

□

Proposition 3.2.4 *Soient* $\mathbf{x}, \mathbf{y} \in I\!R^n$. *Alors,*
$\mathbf{x}$ *et* $\mathbf{y}$ *commutent si et seulement si l'une des trois affirmations suivantes est vérifiée :*

1. $\bar{\mathbf{x}} = 0$,
2. $\bar{\mathbf{y}} = 0$,
3. $\bar{\mathbf{x}}, \bar{\mathbf{y}} \neq 0$ *et* $\bar{\mathbf{x}} = \alpha \bar{\mathbf{y}}$ *avec* $\alpha \neq 0 \in I\!R$.

Preuve :

$\Longrightarrow$:

Supposons que $\mathbf{x}$ et $\mathbf{y}$ partagent la structure de Jordan $\{\mathbf{c_1}, \mathbf{c_2}\}$. Dans le plan engendré par $\mathbf{c_1}$ et $\mathbf{c_2}$ comprenant l'axe des coordonnées x_0, les vecteurs $\mathbf{x}$ et $\mathbf{y}$ peuvent être disposés de plusieurs façons. Si l'un des deux (ou les deux), disons $\mathbf{x}$ pour fixer les idées, se trouve sur l'axe des coordonnées x_0, alors $\bar{\mathbf{x}} = 0$. Si aucun des deux ne se trouve sur cet axe, nous avons $\bar{\mathbf{x}}, \bar{\mathbf{y}} \neq 0$, et en écrivant

$$\mathbf{x} = \lambda_1 \mathbf{c_1} + \lambda_2 \mathbf{c_2} \quad \text{et} \quad \mathbf{y} = \omega_1 \mathbf{c_1} + \omega_2 \mathbf{c_2},$$

nous obtenons

$$\bar{\mathbf{x}} = \lambda_1 \bar{\mathbf{c}}_1 + \lambda_2 \bar{\mathbf{c}}_2 = \lambda_1 \frac{1}{2} \frac{\bar{\mathbf{x}}}{\|\bar{\mathbf{x}}\|} - \lambda_2 \frac{1}{2} \frac{\bar{\mathbf{x}}}{\|\bar{\mathbf{x}}\|} = (\lambda_1 - \lambda_2) \frac{1}{2} \frac{\bar{\mathbf{x}}}{\|\bar{\mathbf{x}}\|}$$

$$= (\lambda_1 - \lambda_2) \frac{\omega_1 - \omega_2}{\omega_1 - \omega_2} \frac{1}{2} \frac{\bar{\mathbf{x}}}{\|\bar{\mathbf{x}}\|} = \alpha(\omega_1 - \omega_2) \frac{1}{2} \frac{\bar{\mathbf{x}}}{\|\bar{\mathbf{x}}\|} = \alpha \bar{\mathbf{y}},$$

où, nécessairement, $\alpha \neq 0$.

$\Longleftarrow$:

Dans les cas $\bar{\mathbf{x}} = 0$ ou $\bar{\mathbf{y}} = 0$ la thèse est vérifiée en utilisant la proposition (3.2.3). Considérons alors le cas $\bar{\mathbf{x}}, \bar{\mathbf{y}} \neq 0$ et $\bar{\mathbf{x}} = \alpha \bar{\mathbf{y}}$ où $\alpha \neq 0 \in I\!R$ (s.p.d.g., $\alpha > 0$).

Nous avons pour $\mathbf{x}$,

$$\mathbf{x} = \lambda_1 \begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \frac{\bar{\mathbf{x}}}{\|\bar{\mathbf{x}}\|} \end{pmatrix} + \lambda_2 \begin{pmatrix} \frac{1}{2} \\ -\frac{1}{2} \frac{\bar{\mathbf{x}}}{\|\bar{\mathbf{x}}\|} \end{pmatrix} = \lambda_1 \mathbf{c_1} + \lambda_2 \mathbf{c_2}$$

et pour $\mathbf{y}$

$$\mathbf{y} = \omega_1 \begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \frac{\bar{\mathbf{y}}}{\|\bar{\mathbf{y}}\|} \end{pmatrix} + \omega_2 \begin{pmatrix} \frac{1}{2} \\ -\frac{1}{2} \frac{\bar{\mathbf{y}}}{\|\bar{\mathbf{y}}\|} \end{pmatrix} = \omega_1 \begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \frac{\bar{\mathbf{x}}}{\|\bar{\mathbf{x}}\|} \end{pmatrix} + \omega_2 \begin{pmatrix} \frac{1}{2} \\ -\frac{1}{2} \frac{\bar{\mathbf{x}}}{\|\bar{\mathbf{x}}\|} \end{pmatrix}$$

$$= \omega_1 \mathbf{c_1} + \omega_2 \mathbf{c_2},$$

qui montre que $\mathbf{x}$ et $\mathbf{y}$ commutent.

□

Théorème 3.2.2 *Pour tout* $\mathbf{x}, \mathbf{y} \in I\!R^n$, *les assertions suivantes sont équivalentes :*

1. *Les opérateurs* $\mathbf{x}$ *et* $\mathbf{y}$ *commutent.*
2. *Arw*$(\mathbf{x})$ *et Arw*$(\mathbf{y})$ *commutent. Donc,* $\forall \mathbf{z} \in I\!R^n$, $\mathbf{x} \circ (\mathbf{y} \circ \mathbf{z}) = \mathbf{y} \circ (\mathbf{x} \circ \mathbf{z})$.
3. $Q_{\mathbf{x}}$ *et* $Q_{\mathbf{y}}$ *commutent.*

Preuve :

1) $\Rightarrow$ 2)

Supposons que $\mathbf{x}$ et $\mathbf{y}$ partagent la structure de Jordan $\{\mathbf{c}_1, \mathbf{c}_2\}$ où, s.p.d.g., $\mathbf{c}_1 = \frac{1}{2}(1; \bar{\mathbf{x}}/\|\bar{\mathbf{x}}\|)$ et $\mathbf{c}_2 = \frac{1}{2}(1; -\bar{\mathbf{x}}/\|\bar{\mathbf{x}}\|)$. Notons, en premier lieu, que $Arw(\mathbf{c}_1)$ et $Arw(\mathbf{c}_2)$ commutent puisque

$$Arw(\mathbf{c}_1)Arw(\mathbf{c}_2) = \frac{1}{4}\begin{pmatrix} 1 & \frac{\bar{\mathbf{x}}^T}{\|\bar{\mathbf{x}}\|} \\ \frac{\bar{\mathbf{x}}}{\|\bar{\mathbf{x}}\|} & I \end{pmatrix}\begin{pmatrix} 1 & -\frac{\bar{\mathbf{x}}^T}{\|\bar{\mathbf{x}}\|} \\ -\frac{\bar{\mathbf{x}}}{\|\bar{\mathbf{x}}\|} & I \end{pmatrix} = \frac{1}{4}\begin{pmatrix} 0 & \mathbf{0}^T \\ \mathbf{0} & I - \frac{\bar{\mathbf{x}}\bar{\mathbf{x}}^T}{\|\bar{\mathbf{x}}\|^2} \end{pmatrix},$$

et

$Arw(\mathbf{c}_1)Arw(\mathbf{c}_2) = (Arw(\mathbf{c}_1)Arw(\mathbf{c}_2))^T = Arw(\mathbf{c}_2)^T Arw(\mathbf{c}_1)^T = Arw(\mathbf{c}_2)Arw(\mathbf{c}_1)$.

Par conséquent,

$$\begin{aligned} Arw(\mathbf{x})Arw(\mathbf{y}) &= Arw(\lambda_1\mathbf{c}_1 + \lambda_2\mathbf{c}_2)Arw(\omega_1\mathbf{c}_1 + \omega_2\mathbf{c}_2) \\ &= [\lambda_1 Arw(\mathbf{c}_1) + \lambda_2 Arw(\mathbf{c}_2)]\,[\omega_1 Arw(\mathbf{c}_1) + \omega_2 Arw(\mathbf{c}_2)] \\ &= \lambda_1\omega_1 Arw^2(\mathbf{c}_1) + \{\lambda_1\omega_2 + \lambda_2\omega_1\}Arw(\mathbf{c}_1)Arw(\mathbf{c}_2) + \lambda_2\omega_2 Arw^2(\mathbf{c}_2) \\ &= [\omega_1 Arw(\mathbf{c}_1) + \omega_2 Arw(\mathbf{c}_2)]\,[\lambda_1 Arw(\mathbf{c}_1) + \lambda_2 Arw(\mathbf{c}_2)] \\ &= Arw(\mathbf{y})Arw(\mathbf{x}). \end{aligned}$$

2) $\Rightarrow$ 3)

Ici encore, nous utilisons l'équivalence entre "commuter" et "posséder les mêmes vecteurs propres".

Par le théorème (3.2.1), $Q_{\mathbf{x}}$ et $Arw(\mathbf{x})$ possèdent les mêmes vecteurs propres. Il en est de même pour $Q_{\mathbf{y}}$ et $Arw(\mathbf{y})$. L'assertion 3) est alors vérifiée puisque 2) revient à dire que $Arw(\mathbf{x})$ et $Arw(\mathbf{y})$ possèdent les mêmes vecteurs propres.

3) $\Rightarrow$ 1)

Supposons que $Q_{\mathbf{x}}$ et $Q_{\mathbf{y}}$ possèdent les mêmes vecteurs propres ; ceci implique que l'ensemble des vecteurs propres de $Q_{\mathbf{x}}$ correspondant aux valeurs propres de multiplicité 1 coïncide avec celui de $Q_{\mathbf{y}}$. Or, ces vecteurs propres forment les structures de Jordan de $\mathbf{x}$ et $\mathbf{y}$, respectivement. D'où, $\mathbf{x}$ et $\mathbf{y}$ commutent.

□

Une conséquence immédiate de ce théorème provient du fait que si $\mathbf{x}$ possède la structure de Jordan $\{\mathbf{c}_1, \mathbf{c}_2\}$, alors par définition, toutes les puissances de $\mathbf{x}$ (si elles sont bien définies) se décomposent dans cette structure de Jordan et donc commutent entre elles. Cette observation mène au corollaire suivant :

Corollaire 3.2.2 *Soit $\mathbf{x} \in I\!R^n$ tel que $\mathbf{x} = \lambda_1\mathbf{c}_1 + \lambda_2\mathbf{c}_2$ et soient p, q des nombres réels arbitraires tels que λ_i^p et λ_i^q soient bien définies pour $i = 1, 2$. Alors,*

1. *$Arw(\mathbf{x}^p)$ et $Arw(\mathbf{x}^q)$ commutent,*
2. *$Q_{\mathbf{x}^p}$ et $Q_{\mathbf{x}^q}$ commutent.*

3.3 Propriétés de $Q_{\mathbf{x}}$

Le théorème qui suit va préparer le terrain pour les chapitres à suivre en présentant les propriétés algébriques fondamentales que vérifie la matrice de représentation quadratique $Q_{\mathbf{x}}$.

Théorème 3.3.1 *Pour tout* $\mathbf{x}, \mathbf{y}, \mathbf{u} \in I\!R^n$, *tels que* $\mathbf{x}$ *est non-singulier,* $\alpha \in I\!R$ *et tout entier* t, *nous avons*

1. $Q_{\alpha\mathbf{y}} = \alpha^2 Q_{\mathbf{y}}$,
2. $Q_{\mathbf{x}}\mathbf{x}^{-1} = \mathbf{x}$ *et donc* $Q_{\mathbf{x}}^{-1}\mathbf{x} = \mathbf{x}^{-1}$,
3. $Q_{\mathbf{y}}\mathbf{e} = \mathbf{y}^2$,
4. $\det(Q_{\mathbf{u}}\mathbf{y}) = \det^2(\mathbf{u})\det(\mathbf{y})$,
5. $Q_{\mathbf{x}^{-1}} = Q_{\mathbf{x}}^{-1}$ *et plus généralement,* $Q_{\mathbf{x}^t} = Q_{\mathbf{x}}^t$.
   Si en plus, $\mathbf{x} \in \mathcal{Q}$ *alors* $Q_{\mathbf{x}^{1/2}} = Q_{\mathbf{x}}^{1/2}$,
6. *Si* $\mathbf{y}$ *est non-singulier,* $(Q_{\mathbf{x}}\mathbf{y})^{-1} = Q_{\mathbf{x}^{-1}}\mathbf{y}^{-1}$, *et plus généralement si* $\mathbf{x}$ *et* $\mathbf{y}$ *commutent, alors* $(Q_{\mathbf{x}}\mathbf{y})^{\alpha} = Q_{\mathbf{x}^{\alpha}}\mathbf{y}^{\alpha}$ *pour des valeurs de* α *telles que* $\mathbf{x}^{\alpha}$ *et* $\mathbf{y}^{\alpha}$ *soient bien définies.*
7. *Lorsqu'en particulier* $\mathbf{x} \succ_{\mathcal{Q}} 0$, *le gradient* $\nabla_{\mathbf{x}}(\ln\det(\mathbf{x})) = 2\mathbf{x}^{-1}$, *et la matrice hessienne* $\nabla_{\mathbf{x}}^2(\ln\det(\mathbf{x})) = -2Q_{\mathbf{x}^{-1}}$,
8. $Q_{Q_{\mathbf{y}}\mathbf{u}} = Q_{\mathbf{y}}Q_{\mathbf{u}}Q_{\mathbf{y}}$,
9. $Q_{\mathbf{y},\mathbf{u}} = \frac{1}{2}(Q_{\mathbf{y}+\mathbf{u}} - Q_{\mathbf{y}} - Q_{\mathbf{u}})$,
10. $Q_{\mathbf{x},\mathbf{x}^{-1}}Q_{\mathbf{x}} = Q_{\mathbf{x}}Q_{\mathbf{x},\mathbf{x}^{-1}} = Arw(\mathbf{x}^2)$

Preuve :
1)

$$\begin{aligned} Q_{\alpha\mathbf{y}} &= 2Arw^2(\alpha\mathbf{y}) - Arw((\alpha\mathbf{y})^2) = 2\alpha^2 Arw^2(\mathbf{y}) - Arw(\alpha^2\mathbf{y}^2) \\ &= \alpha^2(2Arw^2(\mathbf{y}) - Arw(\mathbf{y}^2)) = \alpha^2 Q_{\mathbf{y}}. \end{aligned}$$

2) $Q_{\mathbf{x}}\mathbf{x}^{-1} = 2\mathbf{x} \circ (\mathbf{x} \circ \mathbf{x}^{-1}) - \mathbf{x}^2 \circ \mathbf{x}^{-1} = 2\mathbf{x} \circ \mathbf{e} - \mathbf{x} = \mathbf{x}$.

3) Grâce au fait que $Arw(\mathbf{y})$ et $Q_{\mathbf{y}}$ commutent, nous avons :

$$Q_{\mathbf{y}}\mathbf{e} = Q_{\mathbf{y}}(\mathbf{y} \circ \mathbf{y}^{-1}) = Q_{\mathbf{y}}Arw(\mathbf{y})\mathbf{y}^{-1} = Arw(\mathbf{y})Q_{\mathbf{y}}\mathbf{y}^{-1} = Arw(\mathbf{y})\mathbf{y} = \mathbf{y}^2.$$

4) Posons $\mathbf{z} = Q_{\mathbf{u}}\mathbf{y}$, $\alpha = \mathbf{u}^T\mathbf{y}$ et $\gamma = \det(\mathbf{u})$. Avec ces notations, $\mathbf{z}$ s'écrit $\mathbf{z} = 2\alpha\mathbf{u} - \gamma R\mathbf{y}$ et donc

$$\begin{aligned} z_0^2 &= (2\alpha u_0 - \gamma y_0)^2 = 4\alpha^2 u_0^2 - 4\alpha\gamma u_0 y_0 + \gamma^2 y_0^2 \\ \|\bar{\mathbf{z}}\|^2 &= \|2\alpha\bar{\mathbf{u}} + \gamma\bar{\mathbf{y}}\|^2 = 4\alpha^2\|\bar{\mathbf{u}}\|^2 + 4\alpha\gamma\bar{\mathbf{u}}^T\bar{\mathbf{y}} + \gamma^2\|\bar{\mathbf{y}}\|^2. \end{aligned}$$

Par conséquent, nous obtenons

$$\det(\mathbf{z}) = z_0^2 - \|\bar{\mathbf{z}}\|^2 = 4\alpha^2(u_0^2 - \|\bar{\mathbf{u}}\|^2) - 4\alpha\gamma(u_0 y_0 + \bar{\mathbf{u}}^T\bar{\mathbf{y}}) + \gamma^2(y_0^2 - \|\bar{\mathbf{y}}\|^2)$$

$$= 4\alpha^2\gamma - 4\alpha^2\gamma + \gamma^2(y_0^2 - \|\bar{\mathbf{y}}\|^2) = \det^2(\mathbf{u})\det(\mathbf{y}).$$

5) Notons tout d'abord que $(Q_{\mathbf{x}}R)^2 = \det^2(\mathbf{x})I$. En effet,

$$\begin{aligned}(Q_{\mathbf{x}}R)^2 &= Q_{\mathbf{x}}(RQ_{\mathbf{x}}R) = (2\mathbf{x}\mathbf{x}^T - \det(\mathbf{x})R)(2R\mathbf{x}\mathbf{x}^TR - \det(\mathbf{x})R)\\ &= 4(\mathbf{x}^TR\mathbf{x})\mathbf{x}\mathbf{x}^TR - 4\det(\mathbf{x})\mathbf{x}\mathbf{x}^TR + \det^2(\mathbf{x})I\\ &= \det^2(\mathbf{x})I,\end{aligned}$$

où la dernière égalité vient du fait que $\mathbf{x}^TR\mathbf{x} = \det(\mathbf{x})$. D'autre part,

$$(RQ_{\mathbf{x}})^2 = ((Q_{\mathbf{x}}R)^T)^2 = ((Q_{\mathbf{x}}R)^2)^T = \det^2(\mathbf{x})I.$$

Ensuite, puisque $\mathbf{x}^{-1} = R\mathbf{x}/\det(\mathbf{x})$ nous avons

$$Q_{\mathbf{x}^{-1}} = \frac{1}{\det^2(\mathbf{x})}\begin{pmatrix} \|\mathbf{x}\|^2 & -2x_0\bar{\mathbf{x}}^T \\ -2x_0\bar{\mathbf{x}} & \det(\mathbf{x})I + 2\bar{\mathbf{x}}\bar{\mathbf{x}}^T \end{pmatrix} = \frac{1}{\det^2(\mathbf{x})}RQ_{\mathbf{x}}R.$$

Ainsi,

$$Q_{\mathbf{x}}Q_{\mathbf{x}^{-1}} = \frac{1}{\det^2(\mathbf{x})}(Q_{\mathbf{x}}R)^2 = \frac{1}{\det^2(\mathbf{x})}\det^2(\mathbf{x})I = I,$$

$$Q_{\mathbf{x}^{-1}}Q_{\mathbf{x}} = \frac{1}{\det^2(\mathbf{x})}(RQ_{\mathbf{x}})^2 = \frac{1}{\det^2(\mathbf{x})}\det^2(\mathbf{x})I = I.$$

$\therefore\ Q_{\mathbf{x}^{-1}} = Q_{\mathbf{x}}^{-1}$.

A présent, montrons par récurrence sur $t \geq 0$, que $Q_{\mathbf{x}^t} = Q_{\mathbf{x}}^t$. Pour $t = 0$, nous avons $Q_{\mathbf{x}^0} = Q_{\mathbf{e}} = I = Q_{\mathbf{x}}^0$. Supposons que le résultat soit vrai pour $t \geq 0$ et montrons qu'il l'est encore pour $t+1$. En utilisant les factorisations de Jordan ainsi que le point 3 du théorème (3.2.1) nous obtenons

$$\begin{aligned}Q_{\mathbf{x}}^{t+1} &= Q_{\mathbf{x}}^tQ_{\mathbf{x}} = Q_{\mathbf{x}^t}Q_{\mathbf{x}}\\ &= P\begin{pmatrix} \lambda_1^{2t} & \mathbf{0}^{\mathrm{T}} & 0 \\ \mathbf{0} & \det^t(\mathbf{x})I & \mathbf{0} \\ 0 & \mathbf{0}^T & \lambda_2^{2t} \end{pmatrix}P^TP\begin{pmatrix} \lambda_1^2 & \mathbf{0}^{\mathrm{T}} & 0 \\ \mathbf{0} & \det(\mathbf{x})I & \mathbf{0} \\ 0 & \mathbf{0}^T & \lambda_2^2 \end{pmatrix}P^T\\ &= P\begin{pmatrix} \lambda_1^{2(t+1)} & \mathbf{0}^{\mathrm{T}} & 0 \\ \mathbf{0} & \det^{(t+1)}(\mathbf{x})I & \mathbf{0} \\ 0 & \mathbf{0}^T & \lambda_2^{2(t+1)} \end{pmatrix}P^T = Q_{\mathbf{x}^{(t+1)}},\end{aligned}$$

où $P = (\sqrt{2}\mathbf{c_1}, \widehat{Q}, \sqrt{2}\mathbf{c_2})$ est la matrice orthogonale des vecteurs propres de $Q_{\mathbf{x}}$. Dans le cas $t < 0$ le résultat est également vérifié car

$$Q_{\mathbf{x}}^t = Q_{\mathbf{x}}^{-|t|} = (Q_{\mathbf{x}}^{|t|})^{-1} = (Q_{\mathbf{x}^{|t|}})^{-1} = Q_{\mathbf{x}^{-|t|}} = Q_{\mathbf{x}^t}.$$

$\therefore\ \forall t \in \mathbb{Z},\ Q_{\mathbf{x}}^t = Q_{\mathbf{x}^t}$.

Pour montrer que $Q_{\mathbf{x}^{1/2}} = Q_{\mathbf{x}}^{1/2}$ lorsque $\mathbf{x} \in \mathcal{Q}$, il suffit de montrer que $Q_{\mathbf{x}^{1/2}} Q_{\mathbf{x}^{1/2}} = Q_{\mathbf{x}}$. Mais cela est évident en notant que

$$Q_{\mathbf{x}^{1/2}} = P \begin{pmatrix} \lambda_1 & \mathbf{0}^{\mathrm{T}} & 0 \\ \mathbf{0} & \det^{1/2}(\mathbf{x}) I & \mathbf{0} \\ 0 & \mathbf{0}^T & \lambda_2 \end{pmatrix} P^T.$$

6) Puisque $Q_{\mathbf{x}^{-1}} = RQ_{\mathbf{x}}R/\det^2(\mathbf{x})$ et $\mathbf{y}^{-1} = R\mathbf{y}/\det(\mathbf{y})$, nous avons

$$Q_{\mathbf{x}^{-1}}\mathbf{y}^{-1} = \frac{RQ_{\mathbf{x}}R}{\det^2(\mathbf{x})}\left(\frac{R\mathbf{y}}{\det(\mathbf{y})}\right) = \frac{RQ_{\mathbf{x}}\mathbf{y}}{\det^2(\mathbf{x})\det(\mathbf{y})} = \frac{R(Q_{\mathbf{x}}\mathbf{y})}{\det(Q_{\mathbf{x}}\mathbf{y})} = (Q_{\mathbf{x}}\mathbf{y})^{-1}.$$

A présent, en posant $\mathbf{x} = \lambda_1\mathbf{c}_1 + \lambda_2\mathbf{c}_2$ et $\mathbf{y} = \mu_1\mathbf{c}_1 + \mu_2\mathbf{c}_2$ montrons que pour tout α tel que λ_i^α, μ_i^α soient bien définies pour $i = 1, 2$ nous avons $(Q_{\mathbf{x}}\mathbf{y})^\alpha = Q_{\mathbf{x}^\alpha}\mathbf{y}^\alpha$. Nous avons tout d'abord,

$$\begin{aligned} Q_{\mathbf{x}}\mathbf{y} &= Q_{\mathbf{x}}(\mu_1\mathbf{c}_1 + \mu_2\mathbf{c}_2) = \mu_1 Q_{\mathbf{x}}\mathbf{c}_1 + \mu_2 Q_{\mathbf{x}}\mathbf{c}_2 \\ &= \mu_1\lambda_1^2\mathbf{c}_1 + \mu_2\lambda_2^2\mathbf{c}_2. \end{aligned}$$

Par définition d'une puissance d'un vecteur de $(\mathbb{R}^n, \circ)$, nous obtenons,

$$\begin{aligned} (Q_{\mathbf{x}}\mathbf{y})^\alpha &= (\mu_1\lambda_1^2)^\alpha\mathbf{c}_1 + (\mu_2\lambda_2^2)^\alpha\mathbf{c}_2 = \mu_1^\alpha(\lambda_1^\alpha)^2\mathbf{c}_1 + \mu_2^\alpha(\lambda_2^\alpha)^2\mathbf{c}_2 \\ &= \mu_1^\alpha Q_{\mathbf{x}^\alpha}\mathbf{c}_1 + \mu_2^\alpha Q_{\mathbf{x}^\alpha}\mathbf{c}_2 = Q_{\mathbf{x}^\alpha}(\mu_1^\alpha\mathbf{c}_1 + \mu_2^\alpha\mathbf{c}_2) \\ &= Q_{\mathbf{x}^\alpha}\mathbf{y}^\alpha. \end{aligned}$$

7) Supposons que $\mathbf{x} \succ_{\mathcal{Q}} 0$.

$$\begin{aligned} \nabla_{\mathbf{x}}(\ln\det(\mathbf{x})) &= \nabla_{\mathbf{x}}(\ln(x_0^2 - \textstyle\sum_{i=1}^{n-1} x_i^2)) = \frac{2}{\det(\mathbf{x})}\begin{pmatrix} x_0 \\ -\bar{\mathbf{x}} \end{pmatrix} \\ &= \frac{2}{\det(\mathbf{x})}R\mathbf{x} = 2\mathbf{x}^{-1}. \end{aligned}$$

Ensuite, en remarquant que le gradient de $\det(\mathbf{x})$ est donné par

$$\nabla_{\mathbf{x}}(\det(\mathbf{x})) = \nabla_{\mathbf{x}}\mathbf{x}^T R\mathbf{x} = 2R\mathbf{x}$$

et en partant de $\nabla_{\mathbf{x}}(\ln\det(\mathbf{x})) = \frac{2}{\det(\mathbf{x})}R\mathbf{x}$, nous obtenons pour la matrice hessienne de la fonction $\ln\det(\mathbf{x})$

$$\begin{aligned} \nabla_{\mathbf{x}}^2(\ln\det(\mathbf{x})) &= \frac{2}{\det^2(\mathbf{x})}\left[\det(\mathbf{x})R - 2R\mathbf{x}(R\mathbf{x})^T\right] \\ &= \frac{2}{\det^2(\mathbf{x})}R\left[\det(\mathbf{x})R - 2\mathbf{x}\mathbf{x}^T\right]R \\ &= -\frac{2}{\det^2(\mathbf{x})}RQ_{\mathbf{x}}R = -2Q_{\mathbf{x}^{-1}}. \end{aligned}$$

Notons que l'hypothèse $\mathbf{x} \succ_{\mathcal{Q}} 0$ ($x_0 > \|\bar{\mathbf{x}}\|$) a pour conséquence que toutes les valeurs propres de $Q_{\mathbf{x}}$ (et donc, celles de $Q_{\mathbf{x}^{-1}} = Q_{\mathbf{x}}^{-1}$) sont strictement positives, d'où la matrice hessienne ainsi obtenue est définie négative.

8) En utilisant l'égalité 4) et le fait que $(Q_{\mathbf{y}}R)^2 = \det^2(\mathbf{y})I$ nous avons

$$\begin{aligned} Q_{Q_{\mathbf{y}}\mathbf{u}} &= 2(Q_{\mathbf{y}}\mathbf{u})(Q_{\mathbf{y}}\mathbf{u})^T - \det(Q_{\mathbf{y}}\mathbf{u})R = 2Q_{\mathbf{y}}\mathbf{u}\mathbf{u}^TQ_{\mathbf{y}} - \det^2(\mathbf{y})\det(\mathbf{u})R \\ &= 2Q_{\mathbf{y}}\mathbf{u}\mathbf{u}^TQ_{\mathbf{y}} - \det(\mathbf{u})(Q_{\mathbf{y}}R)^2R = 2Q_{\mathbf{y}}\mathbf{u}\mathbf{u}^TQ_{\mathbf{y}} - \det(\mathbf{u})Q_{\mathbf{y}}RQ_{\mathbf{y}} \\ &= Q_{\mathbf{y}}\{2\mathbf{u}\mathbf{u}^T - \det(\mathbf{u})R\}Q_{\mathbf{y}} = Q_{\mathbf{y}}Q_{\mathbf{u}}Q_{\mathbf{y}}. \end{aligned}$$

9) Par définition de la représentation quadratique, nous pouvons écrire

$$\begin{aligned} & Q_{\mathbf{y}+\mathbf{u}} - Q_{\mathbf{y}} - Q_{\mathbf{u}} = \\ = \; & 2Arw^2(\mathbf{y}+\mathbf{u}) - Arw((\mathbf{y}+\mathbf{u})^2) - 2Arw^2(\mathbf{y}) + Arw(\mathbf{y}^2) - 2Arw^2(\mathbf{u}) \\ & + Arw(\mathbf{u}^2) \\ = \; & 2(Arw(\mathbf{u}) + Arw(\mathbf{y}))^2 - Arw(\mathbf{u}^2 + 2\mathbf{u}\circ\mathbf{y} + \mathbf{y}^2) - 2Arw^2(\mathbf{y}) + Arw(\mathbf{y}^2) \\ & - 2Arw^2(\mathbf{u}) + Arw(\mathbf{u}^2) \\ = \; & 2Arw(\mathbf{u})Arw(\mathbf{y}) + 2Arw(\mathbf{y})Arw(\mathbf{u}) - 2Arw(\mathbf{u}\circ\mathbf{y}) = 2Q_{\mathbf{u},\mathbf{y}}. \end{aligned}$$

10) Nous montrons tout d'abord que : $Arw(\mathbf{x}^{-1})Q_{\mathbf{x}} = Arw(\mathbf{x})$.
Puisque $\mathbf{x}$ et $\mathbf{x}^{-1}$ commutent, en utilisant les théorèmes (3.2.1) et (3.2.2), la matrice $Arw(\mathbf{x})$ commute avec $Q_{\mathbf{x}}$ et $Arw(\mathbf{x}^{-1})$ impliquant que les trois matrices possèdent les mêmes vecteurs propres. Ainsi, pour démontrer cette égalité, il suffit de montrer que la forme de Jordan de $Arw(\mathbf{x})$ est identique au produit des formes de Jordan de $Arw(\mathbf{x}^{-1})$ et $Q_{\mathbf{x}}$. Les formes de Jordan de ces deux dernières matrices sont respectivement :

$$\begin{pmatrix} \lambda_1^{-1} & \mathbf{0}^{\mathrm{T}} & 0 \\ \mathbf{0} & (x^{-1})_0 I & \mathbf{0} \\ 0 & \mathbf{0}^T & \lambda_2^{-1} \end{pmatrix} \quad \text{et} \quad \begin{pmatrix} \lambda_1^2 & \mathbf{0}^{\mathrm{T}} & 0 \\ \mathbf{0} & \lambda_1\lambda_2 I & \mathbf{0} \\ 0 & \mathbf{0}^T & \lambda_2^2 \end{pmatrix}$$

avec $(x^{-1})_0 = (R\mathbf{x})_0/\det(\mathbf{x}) = x_0/\lambda_1\lambda_2$. En effectuant le produit de ces deux matrices, il est clair que nous obtenons la forme de Jordan de $Arw(\mathbf{x})$. L'égalité est alors démontrée.
Notons ensuite que par la définition de $Q_{\mathbf{x},\mathbf{y}}$ et par le fait que $\mathbf{x}$ et $\mathbf{x}^{-1}$ commutent, nous avons :

$$\begin{aligned} Q_{\mathbf{x},\mathbf{x}^{-1}} &= Arw(\mathbf{x})Arw(\mathbf{x}^{-1}) + Arw(\mathbf{x}^{-1})Arw(\mathbf{x}) - Arw(\mathbf{x}\circ\mathbf{x}^{-1}) \\ &= 2Arw(\mathbf{x})Arw(\mathbf{x}^{-1}) - I \end{aligned}$$

et donc,

$$\begin{aligned} Q_{\mathbf{x},\mathbf{x}^{-1}}Q_{\mathbf{x}} &= (2Arw(\mathbf{x})Arw(\mathbf{x}^{-1}) - I)Q_{\mathbf{x}} \\ &= 2Arw(\mathbf{x})Arw(\mathbf{x}) - (2Arw^2(\mathbf{x}) - Arw(\mathbf{x}^2)) \\ &= Arw(\mathbf{x}^2). \end{aligned}$$

De plus, $Q_{\mathbf{x}}Q_{\mathbf{x},\mathbf{x}^{-1}} = (Q_{\mathbf{x},\mathbf{x}^{-1}}Q_{\mathbf{x}})^T = Arw(\mathbf{x}^2)^T = Arw(\mathbf{x}^2)$.

□

Théorème 3.3.2 *Soit* $\mathbf{p} \in I\!R^n$ *non-singulier. Alors, nous avons*

$$Q_{\mathbf{p}}(\mathcal{Q}) = \mathcal{Q}, \quad Q_{\mathbf{p}}(int\,\mathcal{Q}) = int\,\mathcal{Q}.$$

Preuve :
1)$\underline{Q_{\mathbf{p}}(\mathcal{Q}) \subseteq \mathcal{Q}\ (Q_{\mathbf{p}}(\text{int}\,\mathcal{Q}) \subseteq \text{int}\,\mathcal{Q})}$:

Soit $\mathbf{x} \in \mathcal{Q}$ (respectivement, $\mathbf{x} \in \text{int}\,\mathcal{Q}$) et $\mathbf{y} = Q_{\mathbf{p}}\mathbf{x}$. Par le point 4 du théorème (3.3.1), $\det(\mathbf{y}) = (y_0 - \|\bar{\mathbf{y}}\|)(y_0 + \|\bar{\mathbf{y}}\|) = \det^2(\mathbf{p})\det(\mathbf{x}) \geq 0$ (respectivement, $\det(\mathbf{y}) > 0$).

Donc $\begin{cases} (y_0 - \|\bar{\mathbf{y}}\|),\ (y_0 + \|\bar{\mathbf{y}}\|) \geq 0 \text{ (respectivement, } > 0), \text{ ou} \\ (y_0 - \|\bar{\mathbf{y}}\|),\ (y_0 + \|\bar{\mathbf{y}}\|) \leq 0 \text{ (respectivement, } < 0) \end{cases}$

$\Longleftrightarrow \begin{cases} \mathbf{y} \in \mathcal{Q} \text{ (respectivement, int } \mathcal{Q}) \text{ ou} \\ \mathbf{y} \in -\mathcal{Q} \text{ (respectivement, -int } \mathcal{Q}). \end{cases}$

Pour montrer que $\mathbf{y} \in \mathcal{Q}\ (\in \text{int}\,\mathcal{Q})$, il suffit d'obtenir $y_0 \geq 0 (> 0)$. En utilisant le fait que $x_0 \geq (>)\|\bar{\mathbf{x}}\|$ et en appliquant ensuite l'inégalité de Cauchy-Schwarz à $\bar{\mathbf{p}}^T\bar{\mathbf{x}}$ nous trouvons

$$\begin{array}{rcl} y_0 & = & (Q_{\mathbf{p}}\mathbf{x})_0 = (2(\mathbf{p}^T\mathbf{x})\mathbf{p} - \det(\mathbf{p})R\mathbf{x})_0 \\ & = & 2(p_0x_0 + \bar{\mathbf{p}}^T\bar{\mathbf{x}})p_0 - (p_0^2 - \|\bar{\mathbf{p}}\|^2)x_0 \\ & = & x_0p_0^2 + x_0\|\bar{\mathbf{p}}\|^2 + 2p_0(\bar{\mathbf{p}}^T\bar{\mathbf{x}}) \\ (>) & \geq & \|\bar{\mathbf{x}}\|(p_0^2 + \|\bar{\mathbf{p}}\|^2) + 2p_0(\bar{\mathbf{p}}^T\bar{\mathbf{x}}) \\ & \geq & \|\bar{\mathbf{x}}\|(p_0^2 + \|\bar{\mathbf{p}}\|^2) - 2|p_0|\|\bar{\mathbf{p}}\|\|\bar{\mathbf{x}}\| \\ & = & \|\bar{\mathbf{x}}\|(|p_0| - \|\bar{\mathbf{p}}\|)^2 \geq 0. \end{array}$$

$\therefore\ Q_{\mathbf{p}}(\mathcal{Q}) \subseteq \mathcal{Q}\ (Q_{\mathbf{p}}(\text{int}\,\mathcal{Q}) \subseteq \text{int}\,\mathcal{Q})$.

$\underline{2)\mathcal{Q} \subseteq Q_{\mathbf{p}}(\mathcal{Q})\ (\text{int}\,\mathcal{Q} \subseteq Q_{\mathbf{p}}(\text{int}\,\mathcal{Q}))}$:

Puisque $\det(\mathbf{p}^{-1}) = (\det(\mathbf{p}))^{-1} \neq 0$, $\mathbf{p}^{-1}$ est non-singulier et donc, en utilisant 1), nous obtenons $Q_{\mathbf{p}^{-1}}\mathbf{x} \in \mathcal{Q}$ pour tout $\mathbf{x} \in \mathcal{Q}$. Par conséquent, comme pour tout $\mathbf{x}$ nous avons $\mathbf{x} = Q_{\mathbf{p}}(Q_{\mathbf{p}^{-1}}\mathbf{x})$, il suit que $\mathbf{x}$ est l'image par $Q_{\mathbf{p}}$ d'un élément de $\mathcal{Q}$; d'où $\mathbf{x} \in Q_{\mathbf{p}}(\mathcal{Q})$. De façon similaire, nous trouvons $\mathbf{x} \in Q_{\mathbf{p}}(\text{int}\,\mathcal{Q})$ pour tout $\mathbf{x} \in \text{int}\,\mathcal{Q}$.

$\therefore\ \mathcal{Q} \subseteq Q_{\mathbf{p}}(\mathcal{Q})\ (\text{int}\,\mathcal{Q} \subseteq Q_{\mathbf{p}}(\text{int}\,\mathcal{Q}))$.

□

Nous allons à présent prouver un théorème très important qui utilisera de manière explicite le théorème (3.3.2) et dont les résultats seront utilisés dans le chapitre consacré aux méthodes de points intérieurs. Cependant, nous ferons tout d'abord appel à un lemme.

Lemme 3.3.1 *Soient* $\mathbf{x}, \mathbf{y} \in \mathcal{Q}$. *Alors, nous avons*

$\mathbf{x}$ *et* $\mathbf{y}$ *ont les mêmes valeurs propres* $\Leftrightarrow Q_{\mathbf{x}}$ *et* $Q_{\mathbf{y}}$ *sont semblables.*

Preuve : Supposons que les valeurs propres de $\mathbf{x}$ soient $\{\lambda_1, \lambda_2\}$ et celles de $\mathbf{y}$, $\{\omega_1, \omega_2\}$.

$\Leftarrow$:
Si $Q_{\mathbf{x}}$ et $Q_{\mathbf{y}}$ sont semblables, elles possèdent les mêmes valeurs propres (avec les mêmes multiplicités). En utilisant le point 3 du théorème (3.2.1), ceci a pour conséquence que

$$\{\lambda_1^2, \lambda_2^2\} = \{\omega_1^2, \omega_2^2\}$$

De plus, l'hypothèse générale faite sur $\mathbf{x}$ et $\mathbf{y}$ implique que $\lambda_1, \lambda_2, \omega_1, \omega_2 \geq 0$ et donc l'égalité entre ces deux ensembles est équivalente à

$$\{\lambda_1, \lambda_2\} = \{\omega_1, \omega_2\}$$

ce qui montre que $\mathbf{x}$ et $\mathbf{y}$ ont les mêmes valeurs propres .

$\Rightarrow$:
Ecrivons les décompositions de Jordan des matrices $Q_{\mathbf{x}}$ et $Q_{\mathbf{y}}$.

$$Q_{\mathbf{x}} = P \begin{pmatrix} \lambda_1^2 & \mathbf{0}^{\mathrm{T}} & 0 \\ \mathbf{0} & \lambda_1\lambda_2 I & \mathbf{0} \\ 0 & \mathbf{0}^T & \lambda_2^2 \end{pmatrix} P^T \text{ et } Q_{\mathbf{y}} = Q \begin{pmatrix} \omega_1^2 & \mathbf{0}^{\mathrm{T}} & 0 \\ \mathbf{0} & \omega_1\omega_2 I & \mathbf{0} \\ 0 & \mathbf{0}^T & \omega_2^2 \end{pmatrix} Q^T.$$

En appelant les formes de Jordan de $Q_{\mathbf{x}}$ et $Q_{\mathbf{y}}$ respectivement Λ et Ω, nous obtenons $Q_{\mathbf{x}} = P\Lambda P^T$ et $Q_{\mathbf{y}} = Q\Omega Q^T$. Notons que par hypothèse nous avons $\{\lambda_1, \lambda_2\} = \{\omega_1, \omega_2\}$. Deux cas peuvent alors se produire; soit $\lambda_1 = \omega_1$ et $\lambda_2 = \omega_2$, soit $\lambda_1 = \omega_2$ et $\lambda_2 = \omega_1$. Considérons ces deux cas séparement.

$\underline{\lambda_1 = \omega_1 \text{ et } \lambda_2 = \omega_2}$

Dans cette situation, nous pouvons affirmer que $\Lambda = \Omega$ ce qui implique que

$$Q_{\mathbf{y}} = Q\Lambda Q^T = QP^T Q_{\mathbf{x}} PQ^T = (QP^T)Q_{\mathbf{x}}(QP^T)^{-1},$$

où la dernière égalité est obtenue grâce au fait que P et Q sont des matrices orthogonales.

$\underline{\lambda_1 = \omega_2 \text{ et } \lambda_2 = \omega_1}$

Dans ce cas, considérons la matrice de permutation S telle que $S\Omega S^T = \Lambda$; une telle matrice de permutation est orthogonale. Nous obtenons alors,

$$Q_{\mathbf{y}} = Q\Omega Q^T = QS^T S\Omega S^T SQ^T = QS^T \Lambda SQ^T$$

$$= QS^TP^TQ_{\mathbf{x}}PSQ^T = (QS^TP^T)Q_{\mathbf{x}}(QS^TP^T)^{-1}.$$

Nous avons donc montré, dans les deux cas, que $Q_{\mathbf{x}}$ et $Q_{\mathbf{y}}$ sont des matrices semblables.

□

Théorème 3.3.3 *Soient* $\mathbf{x}, \mathbf{z}, \mathbf{p} \in \mathbb{R}^n$ *des vecteurs définis positifs. Définissons* $\widetilde{\mathbf{x}} = Q_{\mathbf{p}}\mathbf{x}$ *et* $\underset{\sim}{\mathbf{z}} = Q_{\mathbf{p}^{-1}}\mathbf{z}$. *Alors,*

1. *les vecteurs* $\mathbf{a} = Q_{\mathbf{x}^{1/2}}\mathbf{z}$ *et* $\mathbf{b} = Q_{\widetilde{\mathbf{x}}^{1/2}}\underset{\sim}{\mathbf{z}}$ *ont le même spectre (i.e., le même ensemble de valeurs propres avec les mêmes multiplicités) ;*
2. *les vecteurs* $Q_{\mathbf{x}^{1/2}}\mathbf{z}$ *et* $Q_{\mathbf{z}^{1/2}}\mathbf{x}$ *ont le même spectre.*

Preuve : En vertu du théorème (3.3.2), les vecteurs $\widetilde{\mathbf{x}}$, $\underset{\sim}{\mathbf{z}}$, $\mathbf{a}$, $\mathbf{b}$, $Q_{\mathbf{x}^{1/2}}\mathbf{z}$ et $Q_{\mathbf{z}^{1/2}}\mathbf{x}$ appartiennent tous à $\operatorname{int} \mathcal{Q}$; cela implique que leurs valeurs propres sont strictement positives. Grâce au lemme (3.3.1), les vecteurs $\mathbf{a}$ et $\mathbf{b}$ possèdent le même spectre si et seulement si les matrices de représentation quadratique qui leur sont associées sont des matrices semblables.

1. En utilisant les parties 5 et 8 du théorème (3.3.1), nous avons

$$Q_{\mathbf{a}} = Q_{Q_{\mathbf{x}^{1/2}}\mathbf{z}} = Q_{\mathbf{x}^{1/2}}Q_{\mathbf{z}}Q_{\mathbf{x}^{1/2}} = Q_{\mathbf{x}}^{1/2}Q_{\mathbf{x}}^{-1}Q_{\mathbf{x}}Q_{\mathbf{z}}Q_{\mathbf{x}}^{1/2} = Q_{\mathbf{x}}^{-1/2}Q_{\mathbf{x}}Q_{\mathbf{z}}Q_{\mathbf{x}}^{1/2}$$

et

$$Q_{\mathbf{b}} = Q_{Q_{\widetilde{\mathbf{x}}^{1/2}}\underset{\sim}{\mathbf{z}}} = Q_{\widetilde{\mathbf{x}}^{1/2}}Q_{\underset{\sim}{\mathbf{z}}}Q_{\widetilde{\mathbf{x}}^{1/2}} = Q_{\widetilde{\mathbf{x}}}^{1/2}Q_{\widetilde{\mathbf{x}}}^{-1}Q_{\widetilde{\mathbf{x}}}Q_{\underset{\sim}{\mathbf{z}}}Q_{\widetilde{\mathbf{x}}}^{1/2} = Q_{\widetilde{\mathbf{x}}}^{-1/2}Q_{\widetilde{\mathbf{x}}}Q_{\underset{\sim}{\mathbf{z}}}Q_{\widetilde{\mathbf{x}}}^{1/2}.$$

Ces deux égalités nous montrent que $Q_{\mathbf{a}}$ est semblable à $Q_{\mathbf{x}}Q_{\mathbf{z}}$ et que $Q_{\mathbf{b}}$ est semblable à $Q_{\widetilde{\mathbf{x}}}Q_{\underset{\sim}{\mathbf{z}}}$. Partant de la définition de $\widetilde{\mathbf{x}}$ et $\underset{\sim}{\mathbf{z}}$, nous obtenons

$$Q_{\widetilde{\mathbf{x}}}Q_{\underset{\sim}{\mathbf{z}}} = Q_{\mathbf{p}}Q_{\mathbf{x}}Q_{\mathbf{p}}Q_{\mathbf{p}^{-1}}Q_{\mathbf{z}}Q_{\mathbf{p}^{-1}} = Q_{\mathbf{p}}Q_{\mathbf{x}}Q_{\mathbf{z}}Q_{\mathbf{p}}^{-1}.$$

Par conséquent, $Q_{\mathbf{a}}$ et $Q_{\mathbf{b}}$ sont semblables, ce qui termine la preuve de la partie 1.

2. Les relations $Q_{Q_{\mathbf{x}^{1/2}}\mathbf{z}} = Q_{\mathbf{x}^{1/2}}Q_{\mathbf{z}}Q_{\mathbf{x}^{1/2}} = Q_{\mathbf{x}}^{-1/2}Q_{\mathbf{x}}Q_{\mathbf{z}}Q_{\mathbf{x}}^{1/2}$ et $Q_{Q_{\mathbf{z}^{1/2}}\mathbf{x}} = Q_{\mathbf{z}^{1/2}}Q_{\mathbf{x}}Q_{\mathbf{z}^{1/2}} = Q_{\mathbf{z}}^{-1/2}Q_{\mathbf{z}}Q_{\mathbf{x}}Q_{\mathbf{z}}^{1/2}$ montrent que $Q_{Q_{\mathbf{x}^{1/2}}\mathbf{z}}$ est semblable à $Q_{\mathbf{x}}Q_{\mathbf{z}}$ et que $Q_{Q_{\mathbf{z}^{1/2}}\mathbf{x}}$ est semblable à $Q_{\mathbf{z}}Q_{\mathbf{x}}$. Mais $Q_{\mathbf{x}}Q_{\mathbf{z}}$ et $Q_{\mathbf{z}}Q_{\mathbf{x}}$ le sont également puisque $Q_{\mathbf{x}}Q_{\mathbf{z}} = Q_{\mathbf{z}}^{-1}Q_{\mathbf{z}}Q_{\mathbf{x}}Q_{\mathbf{z}}$.
Donc, $Q_{Q_{\mathbf{x}^{1/2}}\mathbf{z}}$ et $Q_{Q_{\mathbf{z}^{1/2}}\mathbf{x}}$ sont semblables, et l'application du lemme (3.3.1) nous permet de conclure que les vecteurs $Q_{\mathbf{x}^{1/2}}\mathbf{z}$ et $Q_{\mathbf{z}^{1/2}}\mathbf{x}$ ont le même spectre.

□

Nous terminons cette section en précisant la structure de chaque objet défini jusqu'ici lorsque les vecteurs de l'algèbre SOCP sont partitionnés en blocs.

Définition 3.3.1 *Soient* $\mathbf{x} = (\mathbf{x}_1; \ldots; \mathbf{x}_r)$, $\mathbf{y} = (\mathbf{y}_1; \ldots; \mathbf{y}_r)$, *où* $\mathbf{x}_i, \mathbf{y}_i \in \mathbb{R}^{n_i}$ *pour* $i = 1, \ldots, r$. *Alors,*

1. $\mathbf{x} \circ \mathbf{y} = (\mathbf{x}_1 \circ \mathbf{y}_1; \ldots; \mathbf{x}_r \circ \mathbf{y}_r)$.
2. $Arw(\mathbf{x}) = Arw(\mathbf{x}_1) \oplus \ldots \oplus Arw(\mathbf{x}_r)$.
3. $Q_{\mathbf{x}} = Q_{\mathbf{x}_1} \oplus \ldots \oplus Q_{\mathbf{x}_r}$.
4. $Q_{\mathbf{x},\mathbf{y}} = Q_{\mathbf{x}_1,\mathbf{y}_1} \oplus \ldots \oplus Q_{\mathbf{x}_r,\mathbf{y}_r}$.
5. *Le polynôme caractéristique de* $\mathbf{x}$ *est* $p(\lambda, \mathbf{x}) = \prod_{i=1}^{r} p_i(\lambda, \mathbf{x}_i)$ *et est de degré* $2r$.
6. *La trace de* $\mathbf{x}$ *correspond à l'opposé du coefficient du terme de degré 2r-1 dans le polynôme caractéristique et son déterminant correspond au terme indépendant de ce polynôme.*
7. *Le spectre de* $\mathbf{x}$ *correspond à l'union des spectres de chaque* $\mathbf{x}_i$. *Donc,* $\mathbf{x}$ *possède 2r valeurs propres.*
8. $\|\mathbf{x}\|_F^2 = \sum_{i=1}^{r} \|\mathbf{x}_i\|_F^2$.
9. $\|\mathbf{x}\|_2 = \max_{1 \leq i \leq r} \|\mathbf{x}_i\|_2$.
10. $\mathbf{x}^{-1} = (\mathbf{x}_1^{-1}; \ldots; \mathbf{x}_r^{-1})$ *et plus généralement, pour* $t \in \mathbb{R}$, $\mathbf{x}^t = (\mathbf{x}_1^t; \ldots; \mathbf{x}_r^t)$, *lorsque chaque* $\mathbf{x}_i^t$ *est bien défini.*
11. $\mathbf{x}$ *et* $\mathbf{y}$ *commutent si* $\mathbf{x}_i$ *et* $\mathbf{y}_i$ *commutent pour tout* $i \in \{1, \ldots, r\}$.

Voyons à présent quelles sont les conséquences de ces définitions sur la trace et le déterminant d'un élément partitionné en blocs.

Proposition 3.3.1 *Soit* $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_r) \in \mathbb{R}^n$. *Alors,*

1. $tr(\mathbf{x}) = \sum_{i=1}^{r} tr(\mathbf{x}_i)$.
2. $\det(\mathbf{x}) = \prod_{i=1}^{r} \det(\mathbf{x}_i)$.

Preuve : La preuve consiste à montrer que le coefficient du terme de dégré $2r - 1$ de $p(\lambda, \mathbf{x})$ équivaut à $-\sum_{i=1}^{r} 2x_{i,0}$ et que le terme indépendant de ce même polynôme est $\prod_{i=1}^{r}(x_{i,0}^2 - \|\bar{\mathbf{x}}_i\|^2)$, i.e. ,

$$p(\lambda, \mathbf{x}) = \lambda^{2r} - \sum_{i=1}^{r} 2x_{i,0}\lambda^{2r-1} + \ldots + \prod_{i=1}^{r}(x_{i,0}^2 - \|\bar{\mathbf{x}}_i\|^2). \qquad (3.19)$$

Nous allons procéder par récurrence sur le nombre de blocs r.
Si $r = 1$, nous obtenons immédiatement le résultat puisque, par définition,

$$p(\lambda, \mathbf{x}) = \lambda^2 - 2x_0\lambda + x_0^2 - \|\bar{\mathbf{x}}\|^2.$$

Supposons que (3.19) soit vraie pour $r \geq 1$. Nous aurons alors pour un vecteur $\mathbf{x}$ constitué de $r + 1$ blocs :

$$p(\lambda, \mathbf{x}) = \prod_{i=1}^{r+1} p(\lambda, \mathbf{x}_i) = \prod_{i=1}^{r} p(\lambda, \mathbf{x}_i) p(\lambda, \mathbf{x}_{r+1})$$

$$= (\lambda^{2r} - \sum_{i=1}^{r} 2x_{i,0}\lambda^{2r-1} + \ldots + \prod_{i=1}^{r} (x_{i,0}^2 - \|\bar{\mathbf{x}}_i\|^2))(\lambda^2 - 2x_{r+1,0}\lambda + x_{r+1,0}^2 - \|\bar{\mathbf{x}}_{r+1}\|^2)$$

$$= \lambda^{2(r+1)} - \sum_{i=1}^{r+1} 2x_{i,0}\lambda^{2r+1} + \ldots + \prod_{i=1}^{r+1} (x_{i,0}^2 - \|\bar{\mathbf{x}}_i\|^2),$$

ce qui coïncide bien avec le polynôme (3.19).

$$\therefore \forall r \geq 1, \quad p(\lambda, \mathbf{x}) = \lambda^{2r} - \textstyle\sum_{i=1}^{r} 2x_{i,0}\lambda^{2r-1} + \ldots + \prod_{i=1}^{r} (x_{i,0}^2 - \|\bar{\mathbf{x}}_i\|^2).$$

□

Grâce aux définitions figurant dans (3.3.1) et à la prop (3.3.1), les propositions (3.2.2), (3.2.3), (3.2.4), les théorèmes (3.2.1), (3.2.2), (3.3.1), (3.3.2), (3.3.3) et les corollaires (3.2.1), (3.2.2) se généralisent sans aucune difficulté dans le cas où $\mathbf{x}$ est constitué de plusieurs blocs.

Chapitre 4

Dualité pour SOCP

4.1 Dualité et complémentarité

4.1.1 Dualité faible, semi-forte et forte

Puisque les problèmes SOCP correspondent à des problèmes convexes, il est possible de leur associer une théorie de dualité. Bien qu'une partie importante de cette théorie s'avère être très similaire à celle de la programmation linéaire, elle diffère néanmoins de la programmation linéaire en de nombreux aspects. Nous allons considérer ces similarités et ces différences pour les formes standard primale et duale d'un problème SOCP définies en (1.3).
Comme en programmation linéaire, nous avons pour le saut de dualité :

$$\mathbf{c}^T\mathbf{x} - b^T y = (y^T A + \mathbf{z}^T)\mathbf{x} - \mathbf{x}^T A^T y = \mathbf{x}^T\mathbf{z},$$

et puisque $\mathbf{x} \in \mathcal{Q}$ et $\mathbf{z} \in \mathcal{Q} = \mathcal{Q}^*$ nous avons :

Lemme 4.1.1 (dualité faible)
Soit $\mathbf{x}$ *une solution admissible pour le primal et soit* $(y, \mathbf{z})$ *une solution admissible pour le dual. Alors, le saut de dualité* $\mathbf{c}^T\mathbf{x} - b^T y = \mathbf{z}^T\mathbf{x} \geq 0$.

Le résultat du lemme de dualité faible signifie que si p^* et d^* désignent respectivement la valeur optimale pour le primal et la valeur optimale pour le dual, alors $d^* \leq p^*$. Sous certaines hypothèses, il est possible d'obtenir $p^* = d^*$. Cette propriété s'appelle la dualité forte. En programmation linéaire, la dualité forte est satisfaite dans le cas où le primal ou le dual est admissible. Pour SOCP, nous devrons ajouter une hypothèse plus forte pour avoir la dualité forte. Cette hypothèse est que le primal ou le dual doit être strictement admissible.

Définition 4.1.1 *Un point* $\mathbf{x}$ *(resp.* $(y, \mathbf{z})$*) est dit strictement admissible pour le problème primal (respectivement, dual) si*

$$A\mathbf{x} = b \quad et \quad \mathbf{x} \in int\, \mathcal{Q} \quad (A^T y + \mathbf{z} = \mathbf{c} \ et\ \mathbf{z} \in int\, \mathcal{Q}).$$

Le primal (respectivement, le dual) est non-borné si $p^* = -\infty$ *(respectivement,* $d^* = +\infty$*).*

A partir du lemme de dualité faible, si le primal est non-borné, alors $p^* = d^* = -\infty$. Si le dual est non-borné, alors $d^* = p^* = +\infty$.

Théorème 4.1.1 (dualité semi-forte (a))
Si le problème dual est borné et strictement admissible, alors le problème primal possède une solution et $p^* = d^*$.

Preuve : Puisque $d^* \leq p^*$, tout ce que nous devons montrer est qu'il existe une solution admissible pour le primal $\mathbf{x}^*$ telle que $\mathbf{c}^T\mathbf{x}^* \leq d^*$. Considérons l'ensemble convexe

$$M = \{\mathbf{c} - A^T y \,|\, y \in I\!R^m,\, b^T y \geq d^*\}.$$

1)$M \neq \emptyset$ et $M \cap \text{int}\, \mathcal{Q} = \emptyset$ (en supposant que $b \neq 0$).
En effet, si l'ensemble M était vide, alors nous aurions pour tout $y \in I\!R^m$, $b^T y < d^* \in I\!R$. En particulier pour tout $\lambda > 0$, nous avons $b^T(\lambda b) = \lambda \|b\|^2 < d^*$ ce qui est impossible puisque, par hypothèse, $d^* < +\infty$. Ensuite, supposons que $M \cap \text{int}\, \mathcal{Q} \neq \emptyset$. Alors il existe un $\bar{y} \in I\!R^m$ tel que $b^T\bar{y} \geq d^*$ et $\mathbf{c} - A^T\bar{y} \in \text{int}\, \mathcal{Q}$. Par définition de l'intérieur, il existe un petit voisinage de $\bar{y}$ contenant des points admissibles pour le dual. Puisque $b \neq 0$, il existe des points y dans ce voisinnage tel que $b^T y > b^T\bar{y} \geq d^*$. Ceci est impossible car d^* est la valeur optimale du problème dual. Donc, $M \neq \emptyset$ et $M \cap \text{int}\, \mathcal{Q} = \emptyset$.

2) En appliquant le théorème de séparation d'ensembles convexes aux ensembles convexes M et int $\mathcal{Q}$, nous pouvons affirmer qu'il existe un $\mathbf{x} \in I\!R^n$, $\mathbf{x} \neq 0$ tel que

$$\sup_{\mathbf{z} \in M} \mathbf{x}^T\mathbf{z} \leq \inf_{\mathbf{z} \in \text{int}\, \mathcal{Q}} \mathbf{x}^T\mathbf{z}. \tag{4.1}$$

3)$\mathbf{x} \in \mathcal{Q}$ et $\inf_{\mathbf{z} \in \text{int}\, \mathcal{Q}} \mathbf{x}^T\mathbf{z} = 0$.
Par (4.1), la forme linéaire $\mathbf{x}^T\mathbf{z}$ est bornée inférieurement sur le cône int $\mathcal{Q}$. Cette borne inférieure est plus grande ou égale à zéro. En effet, autrement, il existerait $\hat{\mathbf{z}} \in \text{int}\, \mathcal{Q}$ tel que $\mathbf{x}^T\hat{\mathbf{z}} < 0$. Mais alors $\lim_{\mu \to +\infty} \mathbf{x}^T(\mu\hat{\mathbf{z}}) = -\infty$ ce qui est impossible puisque $\mu\hat{\mathbf{z}} \in \text{int}\, \mathcal{Q}$ pour tout $\mu > 0$ et $M \neq \emptyset$. D'où, $\mathbf{x}^T\mathbf{z} \geq 0$

pour tout $\mathbf{z}$ dans la fermeture de int $\mathcal{Q}$, c'est-à-dire, $\mathcal{Q}$. Nous en concluons que $\mathbf{x} \in \mathcal{Q}^* = \mathcal{Q}$ et que $\inf_{\mathbf{z}\in \text{int } \mathcal{Q}} \mathbf{x}^T\mathbf{z} = 0$ car

$$0 \leq \inf_{\mathbf{z}\in \text{int } \mathcal{Q}} \mathbf{x}^T\mathbf{z} \leq \inf_{\mu>0} \mathbf{x}^T(\mu\widehat{z}) = \inf_{\mu>0} \mu(\mathbf{x}^T\widehat{\mathbf{z}}) = 0,$$

où $\widehat{\mathbf{z}}$ est un élément quelconque de int $\mathcal{Q}$.

4)Il existe $\mu \geq 0$ tq $A\mathbf{x} = \mu b$.
Par 3) et (4.1), nous avons $\sup_{\mathbf{z}\in M} \mathbf{x}^T\mathbf{z} \leq 0$, i.e., en utilisant la définition de M

$$\forall y \in I\!R^m \text{ dans le demi-espace } b^T y \geq d^*, \quad \mathbf{x}^T\mathbf{c} \leq (A\mathbf{x})^T y \qquad (4.2)$$

Nous observons que le problème de programmation linéaire $\min\{(A\mathbf{x})^T y \mid b^T y \geq d^*\}$ est borné inférieurement et par conséquent, par le théorème de dualité en programmation linéaire, son problème dual $\max\{\mu d^* \mid A\mathbf{x} = \mu b,\ \mu \geq 0\}$ est admissible.
$\therefore$ Il existe $\mu \geq 0$ tq $A\mathbf{x} = \mu b$.

5)$\mu > 0$.
Si $\mu = 0$, alors $A\mathbf{x} = 0$ et par (4.2), $\mathbf{x}^T\mathbf{c} \leq 0$. Puisque le dual est strictement admissible par hypothèse, il existe $\hat{y} \in I\!R^m$ tel que $\mathbf{c} - A^T\hat{y} \in \text{int } \mathcal{Q} (= \text{int } \mathcal{Q}^*)$. Puisque $\mathbf{x} \in \mathcal{Q}$, avec $\mathbf{x} \neq 0$, le produit $\mathbf{x}^T(\mathbf{c} - A^T\hat{y})$ est strictement positif. (Autrement, par définition de $\mathcal{Q}^*, \mathbf{x}^T(\mathbf{c} - A^T\hat{y}) = 0$. Soit $\epsilon > 0$ tel que $\mathbf{c} - A^T\hat{y} - \epsilon\mathbf{x} \in \mathcal{Q} (= \mathcal{Q}^*)$. Alors $\mathbf{x}^T(\mathbf{c} - A^T\hat{y} - \epsilon\mathbf{x}) = -\epsilon\|\mathbf{x}\|^2 < 0$ ce qui est impossible car $\mathbf{x} \in \mathcal{Q}$). D'où, $\mathbf{x}^T(\mathbf{c} - A^T\hat{y}) = \mathbf{x}^T\mathbf{c} - (A\mathbf{x})^T\hat{y} = \mathbf{x}^T\mathbf{c} > 0$ qui contredit $\mathbf{x}^T\mathbf{c} \leq 0$.

6)$\mathbf{x}^* = \mathbf{x}/\mu$ est admissible et $\mathbf{c}^T\mathbf{x}^* \leq d^*$.
$\mathbf{x}^*$ est admissible car $A\mathbf{x} = \mu b$ et $\mathbf{x} \in \mathcal{Q}$ qui est un cône. En outre, par (4.2),

$$\forall y \in I\!R^m \text{ dans le demi-espace } b^T y \geq d^*, \quad \mathbf{x}^{*T}\mathbf{c} \leq (A\mathbf{x}^*)^T y = b^T y \leq d^*.$$

Ainsi, $\mathbf{c}^T\mathbf{x}^* \leq d^*$.

7)Le cas $b = 0$.
Lorsque $b = 0$, la valeur optimale du problème dual vaut $d^* = 0$ et $x^* = 0$ est admissible pour le primal et donc finalement $\mathbf{c}^T\mathbf{x}^* \leq d^*$.

□

En dualisant ce résultat et en se servant du fait que le dual du dual est le primal, nous obtenons

Théorème 4.1.2 (dualité semi-forte (b))
Si le problème primal est borné et strictement admissible, alors le problème dual possède une solution et $p^* = d^*$.

Pour illustrer ce dernier résultat, considérons le problème suivant :

$$(P)\left\{\begin{array}{lll}\min & (1,-1,0)\mathbf{x} \\ \text{s.c.} & (0,0,1)\mathbf{x}=1 \\ & \mathbf{x}\succcurlyeq_Q \mathbf{0}\end{array}\right. \quad (D)\left\{\begin{array}{ll}\max & y \\ \text{s.c.} & \begin{pmatrix}0\\0\\1\end{pmatrix} y+\mathbf{z}=\begin{pmatrix}1\\-1\\0\end{pmatrix}. \\ & \mathbf{z}\succcurlyeq_Q \mathbf{0}\end{array}\right.$$

La contrainte linéaire dans (P) exige que $x_2 = 1$; d'où le problème (P) est équivalent au problème

$$\left\{\begin{array}{ll}\min & x_0 - x_1 \\ \text{s.c.} & x_0 \geq \sqrt{x_1^2+1}.\end{array}\right.$$

Puisque sous cette contrainte, $x_0 - x_1 > 0$ et $x_0 - x_1 \to 0$ lorsque $x_0 = \sqrt{x_1^2+1} \to \infty$, nous avons $p^* = \inf(x_0 - x_1) = 0$; le problème (P) ne possède donc pas de solution finie mais la fonction objectif est bornée inférieurement par 0 sur l'ensemble admissible, et (P) est évidemment strictement admissible.
Les contraintes linéaires dans (D) requièrent que $\mathbf{z}^T = (1, -1, -y)$; d'où, (D) est équivalent au problème :

$$\left\{\begin{array}{ll}\max & y \\ \text{s.c.} & 1 \geq \sqrt{1+y^2}.\end{array}\right.$$

$y = 0$ est la seule solution admissible, et par conséquent, la seule solution optimale menant à $d^* = 0$.

En combinant les deux théorèmes de dualité semi-forte précédents nous aboutissons au corollaire de dualité forte suivant :

Corollaire 4.1.1 (dualité forte)
Si les problèmes primal et dual sont strictement admissibles, alors le saut de dualité vaut 0, la valeur optimale commune est finie et chaque problème admet une solution.

4.1.2 Complémentarité

En programmation linéaire, le théorème des écarts complémentaires est une conséquence du fait que si $x \geq 0$ et $z \geq 0$, alors le saut de dualité $x^T z = 0$ si et seulement si $x_i z_i = 0$ pour tout i. Notre but, à présent, est de voir dans le cas d'un prolème SOCP, sous quelle(s) condition(s) nous avons $\mathbf{x}^T\mathbf{z} = 0$ lorsque $\mathbf{x}, \mathbf{z} \succcurlyeq_Q \mathbf{0}$. Le lemme suivant va nous donner une condition équivalente à l'annulation du saut de dualité.

Lemme 4.1.2 (conditions de complémentarité.)
Supposons que $\mathbf{x}, \mathbf{z} \in \mathcal{Q}$, *(i.e.,* $\mathbf{x}_i \succcurlyeq_{\mathcal{Q}_{n_i}} \mathbf{0}$ *et* $\mathbf{z}_i \succcurlyeq_{\mathcal{Q}_{n_i}} \mathbf{0}$, *pour* $i = 1, \ldots, r$*). Alors,* $\mathbf{x}^T\mathbf{z} = 0$ *si et seulement si* $\mathbf{x}_i \circ \mathbf{z}_i = \mathbf{0}$ *pour* $i = 1, \ldots, r$, *ou de manière équivalente,*

1. $\mathbf{x}_i^T\mathbf{z}_i = x_{i0}z_{i0} + \bar{\mathbf{x}}_i{}^T\bar{\mathbf{z}}_i = 0$, $i = 1, \ldots, r$ *et*
2. $x_{i0}\bar{\mathbf{z}}_i + z_{i0}\bar{\mathbf{x}}_i = \mathbf{0}$, $i = 1, \ldots, r$.

Preuve :
$\Rightarrow$:
1) Par hypothèse nous avons $\mathbf{x}_i \in \mathcal{Q}_{n_i}$ et $\mathbf{z}_i \in \mathcal{Q}_{n_i} = \mathcal{Q}_{n_i}^*$ pour $i = 1, \ldots, r$ et donc $\mathbf{x}_i^T\mathbf{z}_i \geq 0$, pour $i = 1, \ldots, r$. Par conséquent, puisque

$$\mathbf{x}^T\mathbf{z} = \sum_{i=1}^{r} \mathbf{x}_i^T\mathbf{z}_i = 0$$

nous obtenons nécessairement pour $i = 1, \ldots, r$, $\mathbf{x}_i^T\mathbf{z}_i = 0$.

2) Pour les i tels que $z_{i0} = 0$ le résultat est établi puisqu'alors $z_{i0} \geq \|\bar{\mathbf{z}}_i\| \Rightarrow z_{i0} = \|\bar{\mathbf{z}}_i\| = 0$.
Soit i tel que $z_{i0} \neq 0$. Puisque par hypothèse $\mathbf{x}_i$, $\mathbf{z}_i \in \mathcal{Q}_{n_i}$, nous pouvons écrire,

$$x_{i0}^2 \geq \|\bar{\mathbf{x}}_i\|^2 \quad z_{i0}^2 \geq \|\bar{\mathbf{z}}_i\|^2.$$

En multipliant la deuxième inégalité de part et d'autre par x_{i0}^2/z_{i0}^2 nous obtenons

$$x_{i0}^2 \geq \|\bar{\mathbf{x}}_i\|^2 \tag{4.3}$$

$$x_{i0}^2 \geq \left(\frac{x_{i0}}{z_{i0}}\right)^2 \|\bar{\mathbf{z}}_i\|^2 \tag{4.4}$$

Par 1) nous avons $x_{i0} = -\frac{\bar{\mathbf{x}}_i{}^T\bar{\mathbf{z}}_i}{z_{i0}}$ ou encore, en multipliant de chaque côté par $-2x_{i0}$,

$$-2x_{i0}^2 = 2\frac{x_{i0}}{z_{i0}}\bar{\mathbf{x}}_i^T\bar{\mathbf{z}}_i. \tag{4.5}$$

Par conséquent, en combinant (4.3), (4.4) et (4.5) nous pouvons écrire

$$0 = x_{i0}^2 + x_{i0}^2 - 2x_{i0}^2 \geq \|\bar{\mathbf{x}}_i\|^2 + \left(\frac{x_{i0}}{z_{i0}}\right)^2 \|\bar{\mathbf{z}}_i\|^2 + 2\frac{x_{i0}}{z_{i0}}\bar{\mathbf{x}}_i^T\bar{\mathbf{z}}_i = \left\|\bar{\mathbf{x}}_i + \frac{x_{i0}}{z_{i0}}\bar{\mathbf{z}}_i\right\|^2.$$

Nous obtenons donc bien $x_{i0}\bar{\mathbf{z}}_i + z_{i0}\bar{\mathbf{x}}_i = \mathbf{0}$.

$\Leftarrow$:
Partant de l'assertion 1) et du fait que $\mathbf{x}^T\mathbf{z} = \sum_{i=1}^{r} \bar{\mathbf{x}}_i^T\bar{\mathbf{z}}_i$, nous obtenons bien $\mathbf{x}^T\mathbf{z} = 0$.

□

Géométriquement, le lemme (4.1.2) nous dit que $\mathbf{x}^T\mathbf{z} = 0$ ($\mathbf{x} \in \mathcal{Q}$ est orthogonal à $\mathbf{z} \in \mathcal{Q}$) si et seulement si pour chaque indice i, nous avons que soit $\mathbf{x}_i$ ou $\mathbf{z}_i$ vaut zero, ou alors ils se trouvent sur des parties opposées du bord du cône $\mathcal{Q}_{n_i}$, i.e., $\mathbf{x}_i$ est la réflexion de $\mathbf{z}_i$ par rapport à l'axe des coordonnées x_0 (point 2 du lemme 4.1.2) en plus d'être orthogonal à $\mathbf{z}_i$ (point 1 du lemme 4.1.2). Le fait que $\mathbf{x}_i$ soit la réflexion de $\mathbf{z}_i$ par rapport à l'axe des coordonnées x_0, se traduit par $\mathbf{x}_i = \alpha_i R\mathbf{z}_i$, avec $\alpha_i > 0$.

En combinant l'admissibilité primale et duale avec les conditions de complémentarité, équivalentes à l'annullation du saut de dualité (lemme 4.1.2), nous aboutissons aux conditions d'optimalité pour une paire primale-duale d'un problème SOCP :

Théorème 4.1.3 (conditions d'optimalité) *Si (P) et (D) sont strictement admissibles, alors* $(\mathbf{x}, \mathbf{z})$ *est une paire de solutions optimales pour (P) et (D) si et seulement si,*

$$\begin{array}{l} A\mathbf{x} = b, \quad \mathbf{x} \in \mathcal{Q} \\ A^T y + \mathbf{z} = \mathbf{c}, \quad \mathbf{z} \in \mathcal{Q} \\ \mathbf{x} \circ \mathbf{z} = \mathbf{0}. \end{array} \tag{4.6}$$

Lorsque $r = 1$, c'est-à-dire lorsque chaque vecteur est constitué d'un seul bloc, le système d'équations (4.6) peut être résolu analytiquement. Si $b = 0$, alors $\mathbf{x} = \mathbf{0}$ et n'importe quelle paire admissible pour le dual $(y, \mathbf{z})$ mènent à un couple de solutions optimales pour (P) et (D). Si $\mathbf{c} \in \text{Span}\,(A^T)$, c'est-à-dire si $A^T y = \mathbf{c}$ pour un certain y, alors $\mathbf{z} = \mathbf{0}$ et cet y joints à un vecteur $\mathbf{x}$ admissible pour le primal, nous donnent également un couple de solutions optimales. Autrement, $\mathbf{x}$ et $\mathbf{z}$ sont non nuls pour toute paire de solutions optimales, et par la dernière équation de (4.6), $\mathbf{z} = \alpha R\mathbf{x}$ où $\alpha = z_0/x_0 > 0$. Pour avoir une expression explicite des solutions $\mathbf{x}$, y, $\mathbf{z}$ dans ce cas précis, nous allons supposer que la matrice $(ARA^T)^{-1}$ est non-singulière.
En remplaçant $\mathbf{z}$ par $\alpha R\mathbf{x}$, nous obtenons à partir des deux premières équations de (4.6)

$$\begin{array}{rcl} A^T y + \alpha R\mathbf{x} = \mathbf{c} & \Rightarrow & ARA^T y + \alpha A\mathbf{x} = AR\mathbf{c} \\ & \Rightarrow & ARA^T y = AR\mathbf{c} - \alpha b \\ & \Rightarrow & y = (ARA^T)^{-1}(AR\mathbf{c} - \alpha b). \end{array}$$

$$\begin{array}{rcl} \mathbf{x} = \frac{1}{\alpha} R\mathbf{z} & \Leftrightarrow & \mathbf{x} = \frac{1}{\alpha} R(\mathbf{c} - A^T y) \\ & \Leftrightarrow & \mathbf{x} = \frac{1}{\alpha} R(\mathbf{c} - A^T[(ARA^T)^{-1}(AR\mathbf{c} - \alpha b)]) \\ & \Leftrightarrow & \mathbf{x} = \frac{1}{\alpha} R(\mathbf{c} - A^T(ARA^T)^{-1}AR\mathbf{c} + \alpha A^T(ARA^T)^{-1}b) \\ & \Leftrightarrow & \mathbf{x} = \frac{1}{\alpha} R\mathbf{c} - \frac{1}{\alpha} RA^T(ARA^T)^{-1}AR\mathbf{c} + RA^T(ARA^T)^{-1}b \\ & \Leftrightarrow & \mathbf{x} = \frac{1}{\alpha} P_R\mathbf{c} + RA^T(ARA^T)^{-1}b. \end{array}$$

où $P_R = R - RA^T(ARA^T)^{-1}AR$.

4.2 Non-dégénérescence et complémentarité stricte

Notre but dans cette section sera d'obtenir des conditions nous garantissant l'unicité d'une paire de solutions optimales pour un problème SOCP et l'inversibilité du Jacobien du système d'équations (4.6) définissant des solutions optimales pour (P) et (D). Nous devrons pour cela aborder les notions de non-dégénérescence et de complémentarité stricte.

4.2.1 Non-dégénérescence

Rappelons que pour un problème SOCP mis sous sa forme primale standard, la région admissible est l'intersection de $\mathcal{Q}$ avec l'ensemble affine $\mathcal{A} = \{\mathbf{x} \in I\!R^n \mid A\mathbf{x} = b\}$. Soit $\mathbf{x}$ un point admissible. La condition de non-dégénérescence en $\mathbf{x}$ revient à imposer que $\mathcal{Q}$ et $\mathcal{A}$ ont une intersection transversalle en ce point. Une intersection transversalle signifie simplement que les espaces tangents en $\mathbf{x}$ pour les deux surfaces $\mathcal{A}$ et $\mathcal{Q}$ engendrent l'espace $I\!R^n$. Cette condition exclut, par exemple, la possibilité pour $\mathcal{A}$ d'être tangent à $\mathcal{Q}$. Pour $\mathcal{A}$, l'espace tangent correspond à $\mathrm{Ker}\, A = \{\mathbf{x} \in I\!R^n \mid A\mathbf{x} = \mathbf{0}\}$ pour tout vecteur dans $\mathcal{A}$. Donc,

Définition 4.2.1 *Soit $\mathcal{T}_{\mathbf{x}}$ l'espace tangent à $\mathcal{Q}$ en $\mathbf{x}$. Alors, un vecteur $\mathbf{x}$ admissible pour le primal est non-dégénéré pour le primal si*

$$\mathcal{T}_{\mathbf{x}} + Ker A = I\!R^n.$$

Sinon, $\mathbf{x}$ est dégénéré pour le primal.

Le vecteur des variables libres y dans le problème dual standard (1.3) peut être éliminé en choisissant une sous-matrice F, $(n-m) \times n$, telle que la matrice $G = (A^T, F^T)$ soit non-singulière et en pré-multipliant l'égalité $A^T y + \mathbf{z} = \mathbf{c}$ par $G^{-1} = (Y^T; Z^T)$, où Y et Z sont de dimensions respectives $n \times m$ et $n \times (n-m)$. En effet, remarquons tout d'abord que puisque $GG^{-1} = I = G^{-1}G$ nous obtenons les conditions suivantes sur les sous-matrices F, Y et Z :

$$G^{-1}G = I \Leftrightarrow \begin{cases} Y^T A^T = I & Y^T F^T = 0 \\ Z^T A^T = 0 & Z^T F^T = I \end{cases} \tag{4.7}$$

$$GG^{-1} = I \Leftrightarrow A^T Y^T + F^T Z^T = I. \tag{4.8}$$

Donc, après une pré-multiplication de $A^T y + \mathbf{z} = \mathbf{c}$ par G^{-1} et en utilisant les identités de (4.7), nous obtenons :

$$\begin{aligned} y + Y^T \mathbf{z} &= Y^T \mathbf{c} \\ Z^T \mathbf{z} &= Z^T \mathbf{c} \end{aligned}$$

Ainsi, dans l'espace des $\mathbf{z}$, l'espace tangent à l'ensemble $\{\mathbf{z} \in I\!R^n \mid Z^T \mathbf{z} = Z^T \mathbf{c}\}$ est égal à Ker (Z^T). Mais (4.7) et (4.8) nous permettent d'obtenir respectivement

Span $A^T \subseteq \text{Ker}\,(Z^T)$ et $\text{Ker}\,(Z^T) \subseteq \text{Span}\,A^T$, i.e., $\text{Ker}\,(Z^T) = \text{Span}\,A^T$. Nous avons alors,

Définition 4.2.2 *Un couple* (y, z) *admissible pour le dual est non-dégénéré pour le dual si*

$$\mathcal{T}_z + \mathit{Span}\,A^T = I\!R^n.$$

Sinon, (y, z) *est dégénéré.*

Avec ces définitions nous possédons une vision intuitive de la notion de non-dégénérescence primale et duale. Cependant, nous aimerions obtenir une façon de caractériser algèbriquement cette notion. Considérons, pour cela, le lemme suivant :

Lemme 4.2.1 *Pour des matrices données* B *et* C *comportant* $n-1$ *lignes, la matrice*

$$M = \begin{pmatrix} B & C \\ 0^T & \mathbf{v}^T \end{pmatrix}$$

est de rang n *pour tout vecteur* $\mathbf{v}$ *non nul si et seulement si* B *est de rang* $n-1$.

Preuve : La preuve consiste à montrer que les n colonnes de M^T sont linéairement indépendantes pour tout vecteur $\mathbf{v}$ non nul $\Leftrightarrow$ les $n-1$ colonnes de B^T sont linéairement indépendantes.

$\underline{\Leftarrow :}$
Soit $s = (s'; s'') \in I\!R^{n-1} \times I\!R = I\!R^n$. Nous devons montrer que

$$\begin{pmatrix} B^T & \mathbf{0} \\ C^T & \mathbf{v} \end{pmatrix} \begin{pmatrix} s' \\ s'' \end{pmatrix} = 0 \Rightarrow s = 0,$$

pour tout vecteur non nul $\mathbf{v}$.
Etant donné que les $n-1$ colonnes de B^T sont linéairement indépendantes, nous déduisons de $B^T s' = 0$ que $s' = 0$ et de $s''\mathbf{v} = 0$ que $s'' = 0$ puisque $\mathbf{v}$ est non nul. D'où, $s = (s'; s'') = 0$.

$\underline{\Rightarrow :}$
Remarque préliminaire : par hypothèse sur les colonnes de M^T, la sous-matrice $\begin{pmatrix} B^T \\ C^T \end{pmatrix}$ doit nécessairement avoir ses $n-1$ colonnes linéairement indépendantes. Cela revient à imposer que $\text{Ker}\,B^T \cap \text{Ker}\,C^T = \{0\}$.

Considérons $s \in I\!R^{n-1}$ tel que $B^T s = 0$ et montrons que $s = 0$. Si nous parvenons à montrer que $C^T s = 0$, nous aurons immédiatement $s = 0$ en

vertu de la remarque faite. Supposons par l'absurde que $C^T s \neq 0$. Alors avec $\mathbf{v} := -C^T s \neq 0$, nous obtenons

$$M^T \begin{pmatrix} s \\ 1 \end{pmatrix} = \begin{pmatrix} B^T & 0 \\ C^T & \mathbf{v} \end{pmatrix} \begin{pmatrix} s \\ 1 \end{pmatrix} = \begin{pmatrix} B^T s \\ C^T s + \mathbf{v} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Or, M^T ayant des colonnes linéairement indépendantes, cela implique que $(s;\ 1) = (0;\ 0)$, ce qui est clairement impossible. Donc, $C^T s = 0$.

□

A partir des formes standard primale et duale, nous allons supposer, s.p.d.g., que tous les blocs $\mathbf{x}_i \in \text{bd}\, \mathcal{Q}_{n_i}$ sont groupés dans un vecteur $\mathbf{x}_B$, tous les blocs $\mathbf{x}_i \in \text{int}\, \mathcal{Q}_{n_i}$ sont groupés dans $\mathbf{x}_I$, et tous les blocs $\mathbf{x}_i = 0$ sont groupés dans $\mathbf{x}_O$. Ainsi, un vecteur $\mathbf{x}$ admissible pour le primal peut être partitionné en trois parties comme suit :

$$\mathbf{x} = (\mathbf{x}_B;\ \mathbf{x}_I;\ \mathbf{x}_O)$$

où $\mathbf{x}_B \in I\!R^{n_B}$, $\mathbf{x}_I \in I\!R^{n_I}$ et $\mathbf{x}_O \in I\!R^{n_O}$; donc, $n_B + n_I + n_O = n$. Nous supposons également que $\mathbf{x}_B$ possède p blocs :

$$\mathbf{x}_B = (\mathbf{x}_1; \ldots; \mathbf{x}_p).$$

Cette partition induit une partition similaire pour la matrice A, i.e.,

$$A = (A_B,\ A_I,\ A_O) \text{ avec } A_B = (A_1, \ldots,\ A_p).$$

Soit $\mathcal{Q}_{n_i} \subseteq I\!R^{n_i}$ un cône du second ordre quelconque du produit cartésien. Pour $\mathbf{x} \in \text{int}\, \mathcal{Q}_{n_i}$, l'espace tangent à $\mathcal{Q}_{n_i}$ en $\mathbf{x}$ vaudra $I\!R^{n_i}$ tout entier car tout vecteur de $I\!R^n$ est tangent à un tel $\mathbf{x}$. Pour $\mathbf{x} = \mathbf{0}$, l'espace tangent est $\{\mathbf{0}\}$. Finalement, pour $\mathbf{x} \in \text{bd}\, \mathcal{Q}_{n_i}$, on peut écrire $\mathbf{x} = \alpha \mathbf{c}'$ où $\mathbf{c}'$ et $\mathbf{c} = R\mathbf{c}'$ forment la structure de Jordan de $\mathbf{x}$. Dans ce cas, l'espace tangent en $\mathbf{x}$ équivaut à l'espace vectoriel à $n-1$ dimensions $\{\mathbf{y} \mid \mathbf{c}^T \mathbf{y} = 0\}$. En particulier, le complémentaire orthogonal de cet espace tangent est la droite $\alpha \mathbf{c}$.
De plus, pour un produit cartésien de cônes du second ordre $\mathcal{Q}_{n_i}$, un espace tangent sera considéré comme un produit cartésien d'espaces tangents $\mathcal{T}_{\mathbf{x}_i}$, ce qui nous donne la caractérisation suivante de la non-dégénérescence primale pour un vecteur partitionné en blocs :

$$(\mathcal{T}_{\mathbf{x}_B} \times \mathcal{T}_{\mathbf{x}_I} \times \mathcal{T}_{\mathbf{x}_O}) + \text{Ker}\ ((A_B,\ A_I,\ A_O)) = I\!R^n. \tag{4.9}$$

Pour deux sous-espaces vectoriels S_1 et S_2 dans $I\!R^n$ nous savons que $(S_1 + S_2)^\perp = S_1^\perp \cap S_2^\perp$. Par conséquent, en prenant le complémentaire orthogonal des deux membres dans (4.9), nous obtenons :

$$((\alpha_1 \mathbf{c}_1) \times \ldots \times (\alpha_p \mathbf{c}_p) \times \{\mathbf{0}\} \times I\!R^{n_O}) \cap \text{Span}\ \left((A_1, \ldots,\ A_p,\ A_I,\ A_O)^T\right) = \{\mathbf{0}\}.$$

Lorsque cette condition est satisfaite, toutes les matrices de la forme

$$H_{\mathcal{Q}} = \begin{pmatrix} A_1 & \cdots & A_p & A_I & A_O \\ \alpha_1 \mathbf{c}_1^T & \cdots & \alpha_p \mathbf{c}_p^T & \mathbf{0}^T & \mathbf{v}^T \end{pmatrix} \tag{4.10}$$

ont des lignes linéairement indépendantes pour $\alpha_1, \ldots, \alpha_p$ et $\mathbf{v}$ non tous nuls.
A présent, souvenons-nous que pour tout $\mathbf{x}_i = \lambda_1^i \mathbf{c'}_i + \lambda_2^i \mathbf{c}_i$, les colonnes de la matrice orthogonale

$$Q_i = (\sqrt{2}\mathbf{c}_i, \widehat{Q}_i, \sqrt{2}\mathbf{c'}_i),$$

où $\widehat{Q}_i \in I\!R^{n_i \times (n_i - 2)}$ est une matrice dont les colonnes sont orthogonales à $\mathbf{c}_i$ et $\mathbf{c'}_i$, sont les vecteurs propres de $Arw(\mathbf{x}_i)$.

Théorème 4.2.1 *Soit* $Q_i = (\sqrt{2}\mathbf{c}_i, \widehat{Q}_i, \sqrt{2}\mathbf{c'}_i) = (\sqrt{2}\mathbf{c}_i, \overline{Q_i})$ *la matrice des vecteurs propres de* $Arw(\mathbf{x}_i)$ *pour chaque bloc* $\mathbf{x}_i = \alpha_i \mathbf{c'}_i \in bd\mathcal{Q}_{n_i}$, $i = 1, \ldots, p$.
Alors, $\mathbf{x} = (\mathbf{x}_1; \cdots; \mathbf{x}_p; \mathbf{x}_I; \mathbf{x}_O)$ *est non-dégénéré pour le primal si et seulement si les lignes de la matrice*

$$(A_1\overline{Q_1}, \cdots, A_p\overline{Q_p}, A_I) \in I\!R^{m \times (n_B + n_I - p)}$$

sont linéairement indépendantes. En particulier, $n_B + n_I - p \geq m$.

Preuve : Considérons la matrice

$$G_Q = Q_1 \oplus \ldots \oplus Q_p \oplus I \oplus I. \tag{4.11}$$

Puisque $Q_i^T \mathbf{c}_i = \frac{\mathbf{e}}{\sqrt{2}}$, $i = 1, \ldots, p$, en post-multipliant H_Q par G_Q nous obtenons

$$H_Q G_Q = \begin{pmatrix} A_1 Q_1 & \cdots & A_p Q_p & A_I & A_O \\ \alpha_1 \mathbf{c}_1^T Q_1 & \cdots & \alpha_p \mathbf{c}_p^T Q_p & \mathbf{0}^T & \mathbf{v}^T \end{pmatrix}$$

$$= \begin{pmatrix} \mathbf{b}_1 & A_1\overline{Q_1} & \cdots & \mathbf{b}_p & A_p\overline{Q_p} & A_I & A_O \\ \gamma_1 & \mathbf{0}^T & \cdots & \gamma_p & \mathbf{0}^T & \mathbf{0}^T & \mathbf{v}^T \end{pmatrix},$$

où $\mathbf{b}_i = \sqrt{2}A_i\mathbf{c}_i$ et $\gamma_i = \frac{\alpha_i}{\sqrt{2}}$ pour $i = 1, \ldots, p$. La matrice $H_Q G_Q$ sera de rang ligne plein si et seulement si c'est la cas pour H_Q puisque G_Q est non-singulière. Or, nous avons vu que, H_Q est de rang ligne plein si et seulement si la propriété de non-dégénérescence primale est satisfaite. Par conséquent, en combinant ces deux observations, nous avons que la propriété de non-dégénérescence primale est satisfaite si et seulement si $H_Q G_Q$ est de rang ligne plein. Enfin, en appliquant le lemme (4.2.1) à $H_Q G_Q$, nous obtenons le résultat attendu.

□

De manière semblable, nous allons considérer tout vecteur $\mathbf{z}$ admissible pour le dual comme étant partitionné en :

$$\mathbf{z} = (\mathbf{z}_B; \mathbf{z}_O; \mathbf{z}_I)$$

avec $\mathbf{z}_B \in I\!R^{m_B}$, $\mathbf{z}_O \in I\!R^{m_O}$ et $\mathbf{z}_I \in I\!R^{m_I}$. Le bloc $\mathbf{z}_B$ est la concaténation de tous les blocs $\mathbf{z}_i$ situés sur bd $\mathcal{Q}_{n_i}$ et il est partitionné en q blocs :

$$\mathbf{z}_B = (\mathbf{z}_1; \cdots; \mathbf{z}_q).$$

Tous les blocs $\mathbf{z}_i$ égaux à 0 sont concaténés pour former $\mathbf{z}_O$, et tous ceux appartenant à int $\mathcal{Q}_{n_i}$ sont concaténés pour former $\mathbf{z}_I$. Nous avons alors, $m_B + m_O + m_I = n$. Nous considérons également pour A un partitionnement identique à celui induit par $\mathbf{z}$:

$$A = \left(\tilde{A}_B,\, \tilde{A}_O,\, \tilde{A}_I\right) \text{ et } \tilde{A}_B = \left(\tilde{A}_1, \cdots,\, \tilde{A}_q\right).$$

(la notation $\tilde{A}$ est utilisée pour distinguer la partition induite par $\mathbf{z}$ de celle induite par $\mathbf{x}$). Pour un tel partitionnement, la non-dégénérescence duale s'exprime par

$$(\mathcal{T}_{\mathbf{z}_B} \times \mathcal{T}_{\mathbf{z}_O} \times \mathcal{T}_{\mathbf{z}_I}) + \text{Span}\left((\tilde{A}_B,\, \tilde{A}_O,\, \tilde{A}_I)^T\right) = I\!R^n.$$

Pour chaque bloc $\mathbf{z}_i \in \text{bd}\, \mathcal{Q}_{n_i}$, $\mathbf{z}_i = \beta_i \mathbf{d}_i'$, où $\mathbf{d}_i = R\mathbf{d}_i'$ et $\mathbf{d}_i'$ forment la structure de Jordan de $\mathbf{z}_i$. En prenant le complémentaire orthogonal de part et d'autre nous obtenons

$$((\beta_1 \mathbf{d}_1) \times \cdots \times (\beta_q \mathbf{d}_q) \times I\!R^{m_O} \times \{\mathbf{0}\}) \cap \text{Ker}\,(\tilde{A}_1, \cdots, \tilde{A}_q,\, \tilde{A}_O,\, \tilde{A}_I) = \{\mathbf{0}\},$$

ce qui signifie que si

$$\beta_1 \tilde{A}_1 \mathbf{d}_1 + \cdots + \beta_q \tilde{A}_q \mathbf{d}_q + \tilde{A}_O \mathbf{v} = \mathbf{0},$$

alors $\beta_i = 0$ pour $i = 1, \ldots, q$, et $\mathbf{v} = \mathbf{0}$. Par conséquent, nous avons

Théorème 4.2.2 *Le couple admissible pour le dual $(y, \mathbf{z})$ où $\mathbf{z} = (\mathbf{z}_1; \cdots; \mathbf{z}_q; \mathbf{z}_O; \mathbf{z}_I)$ est non-dégénéré pour le dual si et seulement si les colonnes de la matrice*

$$(\tilde{A}_1 R\mathbf{z}_1, \cdots,\, \tilde{A}_q R\mathbf{z}_q,\, \tilde{A}_O) \in I\!R^{m \times (q + m_O)}$$

sont linéairement indépendantes. En particulier, $m \geq q + m_O$.

Le théorème qui va suivre nous assure que, comme en programmation linéaire, la non-dégénérescence primale en un point optimal d'un problème SOCP implique l'unicité d'une solution duale et la non-dégénérescence duale implique l'unicité d'une solution primale.

Théorème 4.2.3 *Pour la paire de problèmes SOCP (1.3), les deux assertions suivantes sont vraies :*

1. *Si une solution optimale primale est non-dégénérée, alors la solution optimale duale est unique.*
2. *Si une solution optimale duale est non-dégénérée, alors la solution optimale primale est unique.*

Preuve : Nous allons prouver la contraposée des deux assertions.

1) Soient $(y'; \mathbf{z}')$ et $(y''; \mathbf{z}'')$ deux solutions optimales duales distinctes et soit $\mathbf{x}$ une solution optimale primale. Définissons $(y; \mathbf{z}) = (y' - y''; \mathbf{z}' - \mathbf{z}'')$; d'où, $A^T y + \mathbf{z} = \mathbf{0}$ et y, $\mathbf{z}$ sont non nuls. Supposons que $\mathbf{x}$ est partitionné en $\mathbf{x} = (\mathbf{x}_1; \ldots; \mathbf{x}_p; \mathbf{x}_I; \mathbf{x}_O)$, (avec $\mathbf{x}_i \in \text{bd}\, \mathcal{Q}_{n_i}$, $\mathbf{x}_O = \mathbf{0}$ et $\mathbf{x}_I \succ_{\mathcal{Q}_{n_I}} \mathbf{0}$). Cette partition induit la partition suivante sur A et $\mathbf{z}$:

$$A = (A_1, \ldots, A_p, A_I, A_O) \text{ et } \mathbf{z} = (\mathbf{z}_1; \ldots; \mathbf{z}_p; \mathbf{z}_I; \mathbf{z}_O).$$

En appliquant la condition de complémentarité décrite dans le théorème (4.1.3) aux couples de solutions optimales $(\mathbf{x}, \mathbf{z}')$ et $(\mathbf{x}, \mathbf{z}'')$, nous avons en particulier pour $i = 1, \ldots, p$,

$$\mathbf{z}'_i = \alpha'_i R\mathbf{x}_i \text{ et } \mathbf{z}''_i = \alpha''_i R\mathbf{x}_i.$$

Donc, pour $i = 1, \ldots, p$, $\mathbf{z}_i = \alpha_i R\mathbf{x}_i$, pour certains $\alpha_i \in I\!R$. De plus, puisque $\mathbf{x}_I \succ_{\mathcal{Q}_{n_I}} \mathbf{0}$ et que $\mathbf{z}'_I$ et $\mathbf{z}''_I$ doivent lui être perpendiculaire tout en appartenant à $\mathcal{Q}_{n_I}$, il faut nécessairement que $\mathbf{z}'_I = \mathbf{z}''_I = \mathbf{0}$, ce qui conduit à $\mathbf{z}_I = \mathbf{0}$. De même, puisque $\mathbf{x}_O$ doit être perpendiculaire à $\mathbf{z}'_O$ et $\mathbf{z}''_O$, nous obtenons que $\mathbf{z}'_O$ et $\mathbf{z}''_O$ peuvent être quelconques dans $\mathcal{Q}_{n_O}$, impliquant que $\mathbf{z}_O = \mathbf{z}'_O - \mathbf{z}''_O \in I\!R_{n_O}$. L'égalité $A^T y + \mathbf{z} = \mathbf{0}$ avec $y \neq 0$ a pour conséquence que la matrice

$$\begin{pmatrix} A_1 & \ldots & A_p & A_I & A_O \\ \mathbf{z}_1^T & \ldots & \mathbf{z}_p^T & \mathbf{0}^T & \mathbf{z}_O^T \end{pmatrix}$$

a des lignes linéairement dépendantes. Pour chaque $\mathbf{z}_i$ avec $i = 1, \ldots, p$, nous avons $\mathbf{z}_i = \hat{\alpha}_i \mathbf{c}_i$ pour un certain $\hat{\alpha}_i \in I\!R$ si l'on pose $\mathbf{x}_i = \tilde{\alpha}_i \mathbf{c}'_i$. Remarquons que les $\hat{\alpha}_i$, $i = 1, \ldots, p$ et $\mathbf{z}_O$ sont non tous nul sinon $\mathbf{z}$ serait nul, ce qui est impossible. D'où, cette matrice est de la même forme que H_Q dans (4.10), et puisqu'elle n'est pas de rang ligne plein, le vecteur $\mathbf{x}$ est dégénéré pour le primal.

2) Considérons deux solutions optimales primales distinctes $\mathbf{x}'$ et $\mathbf{x}''$ ainsi que $\mathbf{x} = \mathbf{x}' - \mathbf{x}'' \neq \mathbf{0}$. Soit $(y, \mathbf{z})$ une solution optimale duale. La partition $\mathbf{z} = (\mathbf{z}_1; \ldots; \mathbf{z}_q; \mathbf{z}_O; \mathbf{z}_I)$, (où $\mathbf{z}_i \in \text{bd}\, \mathcal{Q}_{n_i}$, $\mathbf{z}_O = \mathbf{0}$ et $\mathbf{z}_I \succ_{\mathcal{Q}_{m_I}} \mathbf{0}$) induit la partition suivante sur A et $\mathbf{x}$

$$A = (\tilde{A}_1, \ldots, \tilde{A}_q, \tilde{A}_O, \tilde{A}_I) \text{ et } \mathbf{x} = (\mathbf{x}_1; \ldots; \mathbf{x}_q; \mathbf{x}_O; \mathbf{x}_I),$$

où, par la condition de complémentarité, pour $i = 1, \ldots, q$, nous avons $\mathbf{x}_i = \beta_i R\mathbf{z}_i$ et $\beta_i \in I\!R$, $\mathbf{x}_I = \mathbf{0} \in I\!R^{m_I}$ et $\mathbf{x}_O \in I\!R^{m_O}$. Puisque $A\mathbf{x} = \mathbf{0}$, nous pouvons écrire

$$\beta_1 \tilde{A}_1 R\mathbf{z}_1 + \ldots + \beta_q \tilde{A}_q R\mathbf{z}_q + \tilde{A}_O \mathbf{x}_O = \mathbf{0},$$

avec les β_i et $\mathbf{x}_O$ non tous nuls car $\mathbf{x}$ est non nul. Mais cette égalité implique que les colonnes de la matrice $(\tilde{A}_1 R\mathbf{z}_1, \ldots, \tilde{A}_q R\mathbf{z}_q, \tilde{A}_O)$ sont linéairement dépendantes. Par conséquent, en vertu du théorème (4.2.2), $(y, \mathbf{z})$ doit être dégénéré pour le dual.

□

4.2.2 Complémentarité stricte

Le but de cette section est d'étudier la notion de complémentarité stricte pour un couple de solutions optimales (x, z) d'un problème SOCP. En programmation linéaire, une paire optimale (x, z) satisfait la complémentarité stricte si $x_i + z_i > 0$, $\forall i = 1, \ldots, n$. Cette propriété peut être étendue de façon naturelle aux problèmes SOCP.

Définition 4.2.3 *Soient* x *et* $(y, \mathbf{z})$ *des solutions optimales pour le primal et le dual d'un problème SOCP. Alors* $(\mathbf{x}, y, \mathbf{z})$ *(*$(\mathbf{x}, \mathbf{z})$*) satisfait la complémentarité stricte (CS) si* $\mathbf{z} + \mathbf{x} \in$ *int* $\mathcal{Q}$.

La définition signifie que pour chaque bloc i, $\mathbf{x}_i + \mathbf{z}_i \in \operatorname{int} \mathcal{Q}_{n_i}$. Par rapport à son cône du second ordre $\mathcal{Q}_{n_i}$, chaque bloc $\mathbf{x}_i$ se trouve dans l'une des trois situations suivantes : $\mathbf{x}_i$ est soit dans l'intérieur de $\mathcal{Q}_{n_i}$, soit sur la frontière de $\mathcal{Q}_{n_i}$, ou soit égal à $\mathbf{0}$. Pour des solutions optimales x et z, nous allons lister dans le tableau suivant les six manières dont peuvent se disposer les blocs $\mathbf{x}_i$ et $\mathbf{z}_i$ et nous précisons les situations qui vérifient (et qui ne vérifient pas) la CS.

$\mathbf{x}_i$		$\mathbf{z}_i$	CS
$\mathbf{x}_i \in \operatorname{int} \mathcal{Q}_{n_i}$	$\Longrightarrow$	$\mathbf{z}_i = \mathbf{0}$	oui
$\mathbf{x}_i = \mathbf{0}$	$\Longleftarrow$	$\mathbf{z}_i \in \operatorname{int} \mathcal{Q}_{n_i}$	oui
$\mathbf{x}_i \in \operatorname{bd} \mathcal{Q}_{n_i}$		$\mathbf{z}_i \in \operatorname{bd} \mathcal{Q}_{n_i}$	oui
$\mathbf{x}_i \in \operatorname{bd} \mathcal{Q}_{n_i}$		$\mathbf{z}_i = \mathbf{0}$	non
$\mathbf{x}_i = \mathbf{0}$		$\mathbf{z}_i \in \operatorname{bd} \mathcal{Q}_{n_i}$	non
$\mathbf{x}_i = \mathbf{0}$		$\mathbf{z}_i = \mathbf{0}$	non

table 4.1: Situations à l'optimum et complémentarité stricte.

Remarquons que la condition de complémentarité figurant dans le système (4.6) impose que lorsqu'un des deux blocs $\mathbf{x}_i$ ou $\mathbf{z}_i$ est dans $\operatorname{int} \mathcal{Q}_{n_i}$, l'autre doit nécessairement être nul car il doit lui être perpendiculaire et appartenir à $\mathcal{Q}_{n_i}$. Ceci a pour conséquence que $\#\{\mathbf{x}_i \in \operatorname{int} \mathcal{Q}_{n_i}\} \leq \#\{\mathbf{z}_i = \mathbf{0}\}$. En échangeant le rôle des variables x et z, nous obtenons de façon semblable que $\#\{\mathbf{z}_i \in \operatorname{int} \mathcal{Q}_{n_i}\} \leq \#\{\mathbf{x}_i = \mathbf{0}\}$. En d'autres termes, nous avons

$$n_I \leq m_O \text{ et } m_I \leq n_O.$$

Dans la troisième situation nous avons, $x_{i0} = \|\bar{\mathbf{x}}_i\|$, $z_{i0} = \|\bar{\mathbf{z}}_i\|$ ainsi que $\mathbf{x}_i^T \mathbf{z}_i = 0$, (i.e., $x_{i0} z_{i0} = -\bar{\mathbf{x}}_i^T \bar{\mathbf{z}}_i$). Par conséquent,

$$(x_{i0} + z_{i0})^2 - \|\bar{\mathbf{x}}_i + \bar{\mathbf{z}}_i\|^2 = 2x_{i0} z_{i0} - 2\bar{\mathbf{x}}_i^T \bar{\mathbf{z}}_i = 4x_{i0} z_{i0} > 0.$$

D'où, $(x_{i0} + z_{i0}) > \|\bar{\mathbf{x}}_i + \bar{\mathbf{z}}_i\|$ et donc $\mathbf{z}_i + \mathbf{x}_i \in \operatorname{int} \mathcal{Q}$.

Corollaire 4.2.1 *Le couple de solutions optimales* $(\mathbf{x}, \mathbf{z})$ *satisfait la complémentarité stricte si et seulement si pour chaque bloc* i*, soit* $\mathbf{x}_i$ *et* $\mathbf{z}_i$ *sont sur* $bd\,\mathcal{Q}_{n_i}$*, ou si l'un est nul, l'autre est dans* $int\,\mathcal{Q}_{n_i}$.

Par la lecture du tableau précédent et en se rappelant que $\mathbf{x} \in \text{bd}\,\mathcal{Q} \Leftrightarrow \det(\mathbf{x}) = 0$, il est clair qu'à l'optimum (i.e., dans n'importe quelle situation du tableau 4.1) nous avons : $z_{i0}\det(\mathbf{x}_i) = x_{i0}\det(\mathbf{z}_i) = 0$. Donc, la complémentarité stricte est satisfaite si et seulement si, pour chaque bloc, exactement un des deux facteurs x_{i0} et $\det(\mathbf{z}_i)$ est nul et exactement un des deux facteurs z_{i0} et $\det(\mathbf{x}_i)$ est nul. Le corollaire précédent nous permet d'obtenir les inégalités suivantes lorsque la complémentarité stricte est satisfaite à l'optimum : $m_B \leq n_B$, $n_B \leq m_B$, $m_O \leq n_I$ et $n_O \leq m_I$. En combinant toutes les inégalités obtenues nous arrivons à une autre caractérisation de la complémentarité stricte :

$$\text{La CS est satisfaite} \iff m_B = n_B,\ m_I = n_O,\ m_O = n_I \text{ et } p = q.$$

En ajoutant à la complémentarité stricte la non-dégénérescence, et en utilisant les théorèmes (4.2.1) et (4.2.2) nous aboutissons au corollaire suivant

Corollaire 4.2.2 *Si* $(\mathbf{x}, \mathbf{z})$ *est une solution optimale satisfaisant la non-dégénérescence primale et duale ainsi que la CS, alors,*

$$n_B + n_I - p \geq m \geq n_I + p.$$

Ceci est à comparer avec le cas en programmation linéaire où la non-dégénérescence primale et duale et la complémentarité stricte impliquent qu'il existe une solution optimale unique (x, z) telle que x et z ont respectivement m composantes non nulles et m composantes nulles.

Un fait important est que contrairement à la programmation linéaire, la non-dégénérescence primale et duale n'implique pas la CS. Pour illustrer cela, considérons l'exemple suivant :

$$(PCQ)\begin{cases} \min\limits_{(x_1,x_2)\in I\!R^2} & -x_1 \\ \text{s.c.} & q_1(x) := \|x\|^2 \leq 1 \\ & q_2(x) := x^T G x \leq 4 \end{cases}$$

où

$$G = \begin{pmatrix} 4 & 2 \\ 2 & 5 \end{pmatrix} = \begin{pmatrix} 2 & 0 \\ 1 & 2 \end{pmatrix} \cdot \begin{pmatrix} 2 & 1 \\ 0 & 2 \end{pmatrix} = B^T B.$$

(PCQ) correspond à un problème convexe à contraintes quadratiques tel que dom(PCQ)=$I\!R^2$ et vérifiant la condition de Slater avec $x = 0$. La recherche d'une

solution du problème (PCQ) est donc équivalente à la résolution du système de KKT correspondant. Le lagrangien de (PCQ) s'écrit

$$L(x;\,\lambda) = -x_1 + \lambda_1(x_1^2 + x_2^2 - 1) + \lambda_2(4x_1^2 + 4x_1x_2 + 5x_2^2 - 4)$$

Le système de KKT pour (PCQ) est alors :

$$\begin{cases} \nabla_x L(x;\,\lambda) = \begin{pmatrix} 2\lambda_1 x_1 + 8\lambda_2 x_1 + 4\lambda_2 x_2 - 1 \\ 2\lambda_1 x_2 + 4\lambda_2 x_1 + 10\lambda_2 x_2 \end{pmatrix} = 0 \\ \\ \lambda_1.(x_1^2 + x_2^2 - 1) = 0 \\ \lambda_2.(4x_1^2 + 4x_1x_2 + 5x_2^2 - 4) = 0 \\ \\ x_1^2 + x_2^2 \leq 1 \\ 4x_1^2 + 4x_1x_2 + 5x_2^2 \leq 4 \\ \\ \lambda_1,\, \lambda_2 \geq 0. \end{cases}$$

La solution unique du système de KKT s'obtient en considérant le cas $\lambda_1 > 0$ et $\lambda_2 = 0$ (les autres cas donnant lieu à des cas impossibles) ; la solution est alors $(x_1^*;\, x_2^*) = (1;\, 0)$ et $(\lambda_1^*;\, \lambda_2^*) = (0.5;\, 0)$.
Ensuite, essayons d'exprimer (PCQ) comme un problème SOCP sous forme standard.

$$\|x\|^2 \leq 1 \Leftrightarrow \|x\| \leq 1 \Leftrightarrow \|x\| \leq x_0 \wedge x_0 = 1 \Leftrightarrow \mathbf{x} := (x_0;\, x) \succcurlyeq_{\mathcal{Q}_3} \mathbf{0} \wedge x_0 = 1.$$

$$x^T G x \leq 4 \Leftrightarrow \|Bx\|^2 \leq 4 \Leftrightarrow \|Bx\| \leq 2 \Leftrightarrow \begin{cases} \|s\| \leq s_0 \wedge s_0 = 2 \\ Bx = s \end{cases}$$

$$\Leftrightarrow \begin{cases} \mathbf{s} := (s_0; s) \succcurlyeq_{\mathcal{Q}_3} \mathbf{0} \wedge s_0 = 2 \\ Bx - s = 0. \end{cases}$$

Ayant transformé $q_1(x)$ et $q_2(x)$ en des contraintes du cône du second ordre, nous pouvons donner le problème SOCP équivalent à (PCQ) sous les formes primale et duale standard :

$$(P) \begin{cases} \min\limits_{\mathbb{R}^6} -x_1 \\ \\ \left(\begin{array}{cccc:ccc} 1 & 0 & 0 & \vdots & 0 & 0 & 0 \\ 0 & 2 & 1 & \vdots & 0 & -1 & 0 \\ 0 & 0 & 2 & \vdots & 0 & 0 & -1 \\ 0 & 0 & 0 & \vdots & 1 & 0 & 0 \end{array}\right) \begin{pmatrix} \mathbf{x} \\ \mathbf{s} \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 2 \end{pmatrix} \\ \\ \mathbf{x} = (x_0;\, x_1;\, x_2) \succcurlyeq_{\mathcal{Q}_3} \mathbf{0},\ \mathbf{s} = (s_0;\, s_1;\, s_2) \succcurlyeq_{\mathcal{Q}_3} \mathbf{0}. \end{cases}$$

$$(D)\begin{cases} \max\limits_{\mathbb{R}^4} y_1 + 2y_4 \\ \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 1 & 2 & 0 \end{pmatrix} y + \mathbf{z}_\mathbf{x} = \begin{pmatrix} 0 \\ -1 \\ 0 \end{pmatrix} \\ \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \end{pmatrix} y + \mathbf{z}_\mathbf{s} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \\ \mathbf{z}_\mathbf{x} \succcurlyeq_{Q_3} \mathbf{0},\ \mathbf{z}_\mathbf{s} \succcurlyeq_{Q_3} \mathbf{0}. \end{cases}$$

Etant donné l'équivalence entre (PCQ) et (P), la solution de (P) sera $\mathbf{x}^* = (1; 1; 0) \in \mathrm{bd}\, Q_3$, $\mathbf{s}^* = (2; 2; 0) \in \mathrm{bd}\, Q_3$. La solution duale $(y^*, (\mathbf{z}_\mathbf{x}^*; \mathbf{z}_\mathbf{s}^*))$ s'obtient en résolvant le système des contraintes de (D) ainsi qu'en imposant les conditions de complémentarité

$$\mathbf{x}^{*T}\mathbf{z}_\mathbf{x}^* = 0, \quad \mathbf{s}^{*T}\mathbf{z}_\mathbf{s}^* = 0.$$

Nous obtenons alors $y^* = (-1; 0; 0; 0)$, $\mathbf{z}_\mathbf{x}^* = (1; -1; 0) = R\mathbf{x} \in \mathrm{bd}\, Q_3$ et $\mathbf{z}_\mathbf{s}^* = (0; 0; 0)$. Puisque $\mathbf{s}^* \in \mathrm{bd}\, Q_3$ et $\mathbf{z}_\mathbf{s}^* = \mathbf{0}$, le couple optimal $(\mathbf{s}^*, \mathbf{z}_\mathbf{s}^*)$ ne satisfait pas la complémentarité stricte. Par contre, la non-dégénérescence primale et duale est bien satisfaite comme nous allons le vérifier. Pour la non-dégénérescence primale, nous devons nous assurer que les lignes de la matrice H_Q sont linéairement indépendantes. Notons tout d'abord que, puisque $\mathbf{x}^*, \mathbf{s}^* \in \mathrm{bd}\, Q_3$, $\mathbf{x}^* = \alpha_1 \mathbf{c}_1'$, où $\mathbf{c}_1'$ et $\mathbf{c}_1 = R\mathbf{c}_1'$ forment la structure de Jordan de $\mathbf{x}^*$. De même $\mathbf{s}^* = \alpha_2 \mathbf{c}_2'$ où $\mathbf{c}_2'$ et $\mathbf{c}_2 = R\mathbf{c}_2'$ forment la structure de Jordan de $\mathbf{s}^*$. D'où,

$$H_Q = \begin{pmatrix} A_1 & A_2 \\ \alpha_1 \mathbf{c}_1^T & \alpha_2 \mathbf{c}_2^T \end{pmatrix} = \left(\begin{array}{ccc:ccc} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 1 & 0 & -1 & 0 \\ 0 & 0 & 2 & 0 & 0 & -1 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & -1 & 0 & 2 & -2 & 0 \end{array}\right)$$

dont les lignes sont bien linéairement indépendantes.
Pour prouver la non-dégénérescence duale nous devons montrer que les colonnes de $(\tilde{A}_1 R\mathbf{z}_\mathbf{x}^*, \tilde{A}_O)$ sont linéairement indépendantes. La partition induite sur A par $(\mathbf{z}_\mathbf{x}^*; \mathbf{z}_\mathbf{s}^*)$ est identique à celle induite par $(\mathbf{x}^*; \mathbf{s}^*)$; d'où, $\tilde{A}_1 = A_1$ et $\tilde{A}_O = A_2$. Par conséquent,

$$(\tilde{A}_1 R\mathbf{z}_\mathbf{x}^*, \tilde{A}_O) = \left(\begin{array}{c:ccc} 1 & 0 & 0 & 0 \\ 2 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 1 & 0 & 0 \end{array}\right),$$

qui possède bien des colonnes linéairement indépendantes puisque son déterminant vaut 1.

4.3 Non-singularité du jacobien

Si $\mathcal{Q}$ est le produit cartésien de r cônes du second ordre, le système (4.6) s'écrit

$$\begin{array}{ll} A_i^T y + \mathbf{z}_i = \mathbf{c}_i, & i = 1, \ldots, r \\ A_1 \mathbf{x}_1 + \ldots + A_r \mathbf{x}_r = b & \\ Arw(\mathbf{x}_i) Arw(\mathbf{z}_i) \mathbf{e} = \mathbf{0}, & i = 1, \ldots, r. \end{array} \tag{4.12}$$

Notre objectif dans cette section est d'arriver à montrer que lorsque la non-dégénérescence primale et duale ainsi que la CS sont satisfaites à l'optimum, le jacobien du système (4.12) est une matrice non-singulière. Pour cela, nous aurons besoin d'étudier un type particulier de matrice en blocs :

Définition 4.3.1 *Soit*

$$J = \begin{pmatrix} 0 & 0 & B_1^T & I & 0 \\ 0 & 0 & B_2^T & 0 & I \\ B_1 & B_2 & 0 & 0 & 0 \\ V_1 & 0 & 0 & U_1 & 0 \\ 0 & V_2 & 0 & 0 & U_2 \end{pmatrix} \tag{4.13}$$

où les premières, deuxièmes, troisièmes, quatrièmes et cinquièmes lignes et colonnes sont de dimensions respectives m, $n-m$, m, m et $n-m$. Nous dirons que J est une matrice primale-duale par blocs sous forme canonique (en abrégé, matrice PDBC) si

1. $B_1 \in I\!R^{m \times m}$ *est non-singulière,*
2. $V_2 \in I\!R^{(n-m)\times(n-m)}$ *et* $U_1 \in I\!R^{m \times m}$ *sont symétriques et définies positives,*
3. $V_1 \in I\!R^{m \times m}$ *et* $U_2 \in I\!R^{(n-m)\times(n-m)}$ *sont symétriques et semi-définies positives,*
4. V_1 *et* U_1 *commutent, ainsi que* V_2 *et* U_2.

Notre but, dans un premier temps, est de montrer que toute matrice PDBC est non-singulière. Nous devrons pour cela nous servir d'un lemme sur les matrices symétriques semi-définies positives.

Lemme 4.3.1 *Soient A et B deux matrices carrées de dimension n, symétriques et semi-définies positives. Alors,*

1. *les valeurs propres du produit AB sont réelles et non-négatives.*
2. *Si en plus, A et B commutent, le produit AB est une matrice symétrique semi-définie positive.*

Preuve : 1) Observons que le résultat est vrai lorsqu'au moins une des deux matrices (disons A) est définie positive, puisqu'alors AB est semblable à une matrice symétrique semi-définie positive :

$$AB = A^{1/2}A^{1/2}BA^{1/2}A^{-1/2}.$$

Cette remarque nous permet de dire que pour tout $\varepsilon > 0$ le produit $(A+\varepsilon I)B = AB + \varepsilon B$ possède des valeurs propres réelles et non-négatives. Or, il est dit dans [13] que le polynôme caractéristique de $AB + \varepsilon B$ est tel que ses coefficients sont des polynômes en ε ce qui a pour conséquence, par la théorie des fonctions algébriques, que les racines de l'équation caractéristique de $AB + \varepsilon B$ (c'est-à-dire ses valeurs propres) sont des fonctions continues de ε. Pour tout $\varepsilon > 0$, appelons $\lambda_i(\varepsilon)$ la i-ème valeur propre réelle et non-négative de $AB + \varepsilon B$, et λ_i la i-ème valeur propre de AB, $i = 1, \ldots, n$.. Puisque pour tout $\varepsilon > 0$ et pour tout $i = 1, \ldots, n$, $\lambda_i(\varepsilon)$ est réelle et non-négative, nous aurons par continuité que cela reste vrai pour $\lim_{\varepsilon \to 0} \lambda_i(\varepsilon) = \lambda_i(0) = \lambda_i$.

2) Si A et B commutent, le produit AB sera une matrice symétrique puisque

$$(AB)^T = B^T A^T = BA = AB.$$

Le fait que AB soit semi-définie positive se déduit alors du point 1.

□

Proposition 4.3.1 *Toute matrice PDBC est non-singulière.*

Preuve : Rappelons tout d'abord le principe général du processus d'élimination de Gauss appliqué à une matrice $A \in \mathbb{R}^{n \times n}$. Ce processus a pour effet d'introduire des éléments nuls dans la partie sous-diagonale de chaque colonne de A. Cela est réalisé en pré-multipliant $n-1$ fois A par des matrices $M_i \in \mathbb{R}^{n \times n}$, $i = 1, \ldots, n-1$ appelées transformations de Gauss. A la fin du procédé nous avons

$$M_{n-1} \ldots M_2 M_1 A = L$$

où L est une matrice triangulaire supérieure. En réalité, chaque matrice M_i est de la forme

$$M_i = \begin{array}{c} \\ i+1 \end{array} \overset{\overset{i}{\downarrow}}{\begin{pmatrix} 1 & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & \ddots & 0 & \cdots & \cdots & 0 \\ \vdots & \ddots & 1 & 0 & \cdots & 0 \\ \vdots & & -\tau_{i,i+1} & 1 & & 0 \\ \vdots & & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & -\tau_{i,n} & 0 & \cdots & 1 \end{pmatrix}} = I - \tau_i e_i^T,$$

avec $\tau_i^T = (0, \ldots, 0, \tau_{i,i+1}, \ldots, \tau_{i,n-1}, \tau_{i,n})^T \in \mathbb{R}^n$. M_i est telle que, en posant $A^{(i-1)} = M_{i-1}M_{i-2}\ldots M_1 A$ où $A^{(0)} = A$, nous avons $A^{(i)} = M_i A^{(i-1)} =$

$$\begin{array}{c} \\ i+1 \end{array} \overset{\overset{i}{\downarrow}}{\begin{pmatrix} \times & \cdots & \cdots & \cdots & \cdots & \times \\ 0 & \ddots & \times & \cdots & \cdots & \times \\ \vdots & \ddots & \times & \times & \cdots & \times \\ \vdots & & 0 & \times & \cdots & \times \\ \vdots & & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \times & \cdots & \times \end{pmatrix}}.$$

Pour annuler la partie sous-diagonale de la i-ème colonne de $A^{(i-1)}$, les $\tau_{i,j}$, $j = i+1, \ldots, n$ figurant dans M_i doivent être choisis comme étant égaux à $a_{i,j}^{(i-1)}/a_{i,i}^{(i-1)}$, où l'élément $a_{i,i}^{(i-1)}$ est appelé le pivot à l'itération i.
Il est important de noter que les matrices M_i sont non-singulières. En effet, si $M_i = I - \tau_i e_i^T$, alors avec $M_i' := I + \tau_i e_i^T$, nous avons

$$M_i M_i' = (I - \tau_i e_i^T)(I + \tau_i e_i^T) = I - \underbrace{(e_i^T \tau_i)}_{=0} \tau_i e_i^T = I$$

L'égalité $M_i' M_i = I$ s'obtient de la même manière et permet de conclure que $M_i^{-1} = M_i'$.

Pour démontrer le résultat annoncé, considérons une matrice PDBC quelconque J. Nous allons appliquer le processus d'élimination de Gauss à J dont les éléments sont des blocs. Tout d'abord, permutons les colonnes 2 et 5 ainsi que les lignes 1 et 3 pour former J'. Pour annuler la partie sous-diagonale de la première colonne de J', nous prémultiplions $J' = J'^{(0)}$ par la transformation de Gauss M_1 afin d'obtenir $J'^{(1)}$

$$M_1 = \begin{pmatrix} I & 0 & 0 & 0 & 0 \\ 0 & I & 0 & 0 & 0 \\ 0 & 0 & I & 0 & 0 \\ -V_1 B_1^{-1} & 0 & 0 & I & 0 \\ 0 & 0 & 0 & 0 & I \end{pmatrix} \Rightarrow J'^{(1)} = \begin{pmatrix} B_1 & 0 & 0 & 0 & B_2 \\ 0 & I & B_2^T & 0 & 0 \\ 0 & 0 & B_1^T & I & 0 \\ 0 & 0 & 0 & U_1 & -V_1 B_1^{-1} B_2 \\ 0 & U_2 & 0 & 0 & V_2 \end{pmatrix}$$

En effectuant encore trois pré-multiplications par des transformations de Gauss, nous aboutissons à

$$J'^{(4)} = \begin{pmatrix} B_1 & 0 & 0 & 0 & B_2 \\ 0 & I & B_2^T & 0 & 0 \\ 0 & 0 & B_1^T & I & 0 \\ 0 & 0 & 0 & U_1 & -V_1 B_1^{-1} B_2 \\ 0 & 0 & 0 & 0 & C \end{pmatrix},$$

où

$$C = V_2(I + V_2^{-1} U_2 B_2^T B_1^{-T} U_1^{-1} V_1 B_1^{-1} B_2).$$

Nous observons que $J'^{(4)}$ est une matrice triangulaire supérieure par blocs qui a été obtenue à partir de J par pré et post-multiplications de matrices non-singulières (matrices de permutation et transformations de Gauss). Par conséquent, J sera non-singulière si et seulement si $J'^{(4)}$ est non-singulière, qui se produit si et seulement si C est non-singulière puisque B_1 et U_1 le sont.
Puisque V_2 est non-singulière et que V_2 commute avec U_2, nous avons que V_2^{-1} commute avec U_2. Ainsi, en appliquant le lemme (4.3.1) avec $A \hookleftarrow V_2^{-1}$ et $B \hookleftarrow U_2$, nous obtenons que $V_2^{-1}U_2$ est une matrice symétrique semi-définie positive. Similairement, puisque U_1 est non-singulière et que V_1 commute avec U_1, nous avons que $B_2^T B_1^{-T} U_1^{-1} V_1 B_1^{-1} B_2$ est symétrique et semi-définie positive. Ensuite, avec $A \hookleftarrow V_2^{-1}U_2$ et $B \hookleftarrow B_2^T B_1^{-T} U_1^{-1} V_1 B_1^{-1} B_2$, le point 1 du lemme (4.3.1) nous permet d'affirmer que le produit $V_2^{-1} U_2 B_2^T B_1^{-T} U_1^{-1} V_1 B_1^{-1} B_2$ possède des valeurs propres réelles non-négatives. Par conséquent, en ajoutant l'identité à cette matrice produit, nous formons une matrice dont les valeurs propres sont strictement positives, qui devra forcément être non-singulière. Finalement, puisque V_2 est non-singulière, C l'est également, ce qui démontre le résultat annoncé.

□

Le Jacobien du système (4.12) est la matrice

$$J_Q = \left(\begin{array}{ccccc|c|ccccc} 0 & \cdots & 0 & 0 & 0 & A_1^T & I & & & & \\ \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & & \ddots & & & \\ 0 & \cdots & 0 & 0 & 0 & A_p^T & & & I & & \\ 0 & \cdots & 0 & 0 & 0 & A_I^T & & & & I & \\ 0 & \cdots & 0 & 0 & 0 & A_O^T & & & & & I \\ \hline A_1 & \cdots & A_p & A_I & A_O & 0 & 0 & \cdots & 0 & 0 & 0 \\ \hline Z_1 & & & & & 0 & X_1 & & & & \\ & \ddots & & & & \vdots & & \ddots & & & \\ & & Z_p & & & 0 & & & X_p & & \\ & & & Z_O & & 0 & & & & X_I & \\ & & & & Z_I & 0 & & & & & X_O \end{array}\right) \tag{4.14}$$

où $X_i = Arw(\mathbf{x}_i)$, $Z_i = Arw(\mathbf{z}_i)$, $i = 1, \ldots, p$, $X_I = Arw(\mathbf{x}_I)$, $Z_I = Arw(\mathbf{z}_I)$, $X_O = Arw(\mathbf{x}_O)$ et $Z_O = Arw(\mathbf{z}_O)$. Remarquons que $X_O = 0$, $Z_O = 0$ et

X_I et Z_I sont symétriques définies positives en vertu de la proposition (1.2.1). Pour montrer que J_Q est non-singulière lorsque la non-dégénérescence primale et duale et la CS sont vérifiées nous allons la transformer en une matrice PDBC. Par les conditions de complémentarité, les solutions optimales $\mathbf{x}$ et $\mathbf{z}$ commutent (cfr. tableau 4.1). D'où, par le théorème (3.2.2), $Arw(\mathbf{x})$ et $Arw(\mathbf{z})$ commutent et partagent donc les mêmes vecteurs propres. De plus, la seule possibilité pour que dans le couple optimal $(\mathbf{x}_i, \mathbf{z}_i)$ ni $\mathbf{x}_i$, ni $\mathbf{z}_i$ ne soit nul est que ces deux vecteurs se trouvent sur bd $\mathcal{Q}_{n_i}$. Dans ce cas, il existe une structure de Jordan $\{\mathbf{c}_i', \mathbf{c}_i = R\mathbf{c}_i'\}$, telle que $\mathbf{x}_i = \alpha_i \mathbf{c}_i'$ et $\mathbf{z}_i = \beta_i \mathbf{c}_i$ avec $\alpha_i = x_{i0} + \|\bar{\mathbf{x}}_i\| = 2x_{i0} > 0$ et $\beta_i = z_{i0} + \|\bar{\mathbf{z}}_i\| = 2z_{i0} > 0$. En posant comme auparavant $Q_i = (\sqrt{2}\mathbf{c}_i', \widehat{Q}_i, \sqrt{2}\mathbf{c}_i)$, la matrice des vecteurs propres communs à $Arw(\mathbf{x}_i)$ et $Arw(\mathbf{z}_i)$ pour chaque bloc sur bd $\mathcal{Q}_{n_i}$, nous avons

$$Q_i^T Arw(\mathbf{x}_i) Q_i = \begin{pmatrix} 2x_{i0} & \mathbf{0}^T & 0 \\ \mathbf{0} & x_{i0} I & \mathbf{0} \\ 0 & \mathbf{0}^T & 0 \end{pmatrix}$$

et

$$Q_i^T Arw(\mathbf{z}_i) Q_i = \begin{pmatrix} 0 & \mathbf{0}^T & 0 \\ \mathbf{0} & z_{i0} I & \mathbf{0} \\ 0 & \mathbf{0}^T & 2z_{i0} \end{pmatrix}.$$

Théorème 4.3.1 *Lorsque, pour un problème SOCP, la non-dégénérescence primale et duale ainsi que la complémentarité stricte sont satisfaites à l'optimum, cela implique que la matrice jacobienne de (4.12) est non-singulière à cet optimum.*

Preuve : Définissons la matrice diagonale par blocs P_Q comme

$$P_Q = (Q_1 \oplus \cdots \oplus Q_p \oplus I \oplus I) \oplus I \oplus (Q_1 \oplus \cdots Q_p \oplus I \oplus I) \tag{4.15}$$

où Q_i correspond à la matrice des vecteurs propres de $Arw(\mathbf{x}_i)$ pour des blocs $\mathbf{x}_i$ sur bd $\mathcal{Q}_{n_i}$. Considérons ensuite la matrice $P_Q^T J_Q P_Q$. Le résultat de cette multiplication matricielle donne $P_Q^T J_Q P_Q =$

$$\left(\begin{array}{ccccc|c|ccccc}
0 & \cdots & 0 & 0 & 0 & Q_1^T A_1^T & I & & & & \\
\vdots & \ddots & \vdots & \vdots & \vdots & \vdots & & \ddots & & & \\
0 & \cdots & 0 & 0 & 0 & Q_p^T A_p^T & & & I & & \\
0 & \cdots & 0 & 0 & 0 & A_I^T & & & & I & \\
0 & \cdots & 0 & 0 & 0 & A_O^T & & & & & I \\
\hline
A_1 Q_1 & \cdots & A_p Q_p & A_I & A_O & 0 & 0 & \cdots & 0 & 0 & 0 \\
\hline
Q_1^T Z_1 Q_1 & & & & & 0 & Q_1^T X_1 Q_1 & & & & \\
& \ddots & & & & \vdots & & \ddots & & & \\
& & Q_p^T Z_p Q_p & & & 0 & & & Q_p^T X_p Q_p & & \\
& & & Z_O & & 0 & & & & X_I & \\
& & & & Z_I & 0 & & & & & X_O
\end{array}\right)$$

Rappelons que la complémentarité stricte implique que $n_B = m_B$ (d'où, $A_i = \tilde{A}_i$, $i = 1, \ldots, p$) et $m_O = n_I$ (d'où, $\tilde{A}_O = A_I$). La non-dégénérescence primale et duale combinée à la complémentarité stricte impliquent alors que

1. la matrice $A' = (A_1\overline{Q_1}, \ldots, A_p\overline{Q_p}, A_I)$ à des lignes linéairement indépendantes et $m \leq n_B + n_I - p$;
2. la matrice $A'' = (A_1\mathbf{c}_1, \ldots, A_p\mathbf{c}_p, A_I)$, où les $\mathbf{c}_i$ sont comme dans le théorème (4.2.2), a des colonnes linéairement indépendantes et donc $m \geq p + n_I$.

Par conséquent, il est possible de prendre les $p+n_I$ colonnes de A'' avec $m-p-n_I$ colonnes de A' afin de construire une matrice $m \times m$ non-singulière B_1. Les colonnes restantes de $(A_1Q_1, \cdots, A_pQ_p, A_I, A_O)$ (au nombre de $n-m$) forment B_2. Une fois B_1 et B_2 construites, les matrices U_1, U_2, V_1 et V_2 sont établies. Plus précisément, les blocs constituant V_2 proviennent des colonnes non nulles des blocs $Q_i^T Z_i Q_i$ ou de Z_I ; V_2 est donc bien définie positive. De même, les blocs constituant U_1 proviennent des colonnes non nulles des blocs $Q_i^T X_i Q_i$ ou de X_I et donc rendent U_1 définie positive. Les colonnes restantes des blocs $Q_i^T X_i Q_i$ et X_O constitueront U_2 et celles de $Q_i^T Z_i Q_i^T$ et Z_O constitueront V_1. D'où, modulo les permutations adéquates, $P_Q^T J_Q P_Q$ est une matrice PDBC, ce qui implique la non-singularité de J_Q.

□

Chapitre 5

Méthodes de points intérieurs

5.1 Introduction

Ce chapitre sera consacré à l'étude de méthodes de points intérieurs de type primal-dual pour la résolution du problème SOCP (1.3).
En programmation linéaire, on distingue généralement deux grandes classes d'algorithmes de points intérieurs. Une de ces deux classes regroupe des méthodes uniquement primales ou duales. Elle inclut l'algorithme de base de Karmarkar (1984) ainsi que beaucoup d'autres algorithmes développés quelques années plus tard après la publication des travaux de Karmarkar. La seconde classe est constituée par les méthodes du type primal-dual développées principalement par M.Kojima, S.Mizuno, A.Yoshise, R.D.C.Monteiro et I.Adler. Le principe général de ces méthodes consiste à appliquer la méthode de Newton au système d'équations $Ax = b, A^T y + z = c$ et $x_i z_i = \mu$ afin d'obtenir une série de directions de Newton menant à l'optimum. Des études empiriques approfondies ont montré qu'en général les méthodes basées sur cette approche primale-duale donnent lieu à des résultats plus favorables d'un point vue numérique par rapport aux méthodes uniquement primales ou duales.
Un de nos objectifs majeurs est de pouvoir étendre ces méthodes primales-duales de la programmation linéaire vers SOCP. Bien que la généralisation naturelle de $x_i z_i = \mu$ pour SOCP soit $\mathbf{x}_i \circ \mathbf{z}_i = \mu \mathbf{e}$, le reste ne notre extension vers SOCP n'aura rien de trivial. Une raison cruciale à cela est que la plupart des méthodes du type primal-dual en programmation linéaire se basent sur le fait trés appréciable que les matrices Diag(x) et Diag(z) commutent. Or, dans SOCP nous perdons cette commodité puisque l'analogue de Diag(x) (=Arw(x)) ne commute pas, en général, avec Arw(z), l'analogue de Diag(z). En réalité, ce problème survient également en programmation semi-définie et plus généralement dans tous les problèmes d'optimisation sur des cônes symétriques.[1] Afin de pallier à ce pro-

[1] Un cône $\mathcal{K}$ est dit symétrique si il est auto-dual et homogène. La propriété d'homogénéité signifie que $\forall x, y \in \text{int}\,\mathcal{K}$ il doit exister une transformation linéaire T telle que $T(x) = y$ et $T(\mathcal{K}) = \mathcal{K}$. Le cône du second ordre est un cône symétrique particulier. En effet, pour deux vecteurs x et y dans int $\mathcal{Q}$, on peut toujours envoyer x sur y en utilisant uniquement des

blème de non-commutativité, les chercheurs ont proposé plusieurs méthodes du type primal-dual différentes. Il y a eu tout d'abord la classe de méthodes dites de Nesterov-Todd (méthode NT, 1997) se basant sur des algorithmes de complexité polynomiale qui ont été généralisés à l'ensemble des cônes symétriques. Ensuite est apparue la classe des méthodes XZ, ZX et $XZ+ZX$ développées principalement par Helmberg, Kojima et Monteiro (1997,1998). Dans le même temps, Monteiro et Zhang (1998) ont développé la famille de Monteiro-Zhang incluant toutes les méthodes citées juste avant comme cas particuliers. Concernant les résultats les plus récents, signalons que Faybusovich a étendu les méthodes NT, XZ, ZX à l'ensemble des cônes symétriques (1998). Schmieta et Alizadeh ont fait de même concernant la famille de Monteiro-Zhang toute entière (1999-2001).

5.2 Barrière logarithmique

Pour suivre le même cheminement qu'en programmation linéaire, nous allons partir du problème SOCP standard (1.3) sous forme primale (problème (P)), construire à partir de celui un problème à barrière logarithmique (P_μ) et appliquer la méthode de Newton au système de KKT de (P_μ).

Considérons la fonction

$$\begin{aligned} \hat{I}: \quad & \mathcal{Q}_n \subset I\!R^n \to I\!R \cup \{+\infty\} \\ & \mathbf{x} \rightsquigarrow \hat{I}(\mathbf{x}) := -\ln\det(\mathbf{x}) = -\ln(x_0^2 - \|\bar{\mathbf{x}}\|^2), \end{aligned}$$

tel que $\mathrm{dom}(\hat{I}) = \mathrm{int}\,\mathcal{Q}_n$. Nous vérifions aisément que $\hat{I}$ est une fonction barrière convexe pour $\mathcal{Q}$ (sous-entendu, $\mathcal{Q}_n$) :

1. D'après sa définition, il est clair que $\hat{I}$ est continue sur $\mathrm{int}\,\mathcal{Q}$. Nous avons montré dans le théorème (3.3.1, point 7) que lorsque $\mathbf{x} \in \mathrm{int}\,\mathcal{Q}$, nous avons

   $$\nabla_{\mathbf{x}}(\ln\det(\mathbf{x})) = 2\mathbf{x}^{-1} \text{ et } \nabla^2_{\mathbf{x}}(\ln\det(\mathbf{x})) = -2Q_{\mathbf{x}^{-1}}.$$

   Puisque $\lambda_1^{-1} = 1/(x_0 + \|\bar{\mathbf{x}}\|)$ et $\lambda_2^{-1} = 1/(x_0 - \|\bar{\mathbf{x}}\|)$ sont les valeurs propres de $\mathbf{x}^{-1}$, $\nabla_{\mathbf{x}}(\ln\det(\mathbf{x}))$ sera continu sur le domaine de $\hat{I}$. En outre, le théorème (3.2.1, point 3) et le fait que $\lambda_{1,2}^{-1} > 0$ sur $\mathrm{int}\,\mathcal{Q}$ nous assurent que $Q_{\mathbf{x}^{-1}}$ est définie positive sur le domaine de $\hat{I}$.

   $\Rightarrow \hat{I}$ est strictement convexe et de classe C^1 sur son domaine.

2. $\mathrm{dom}(\hat{I}) = \{\mathbf{x} \in I\!R^n \mid \mathbf{x} \in \mathrm{int}\,\mathcal{Q}_n\} = \{\mathbf{x} \in I\!R^n \mid x_0 > \|\bar{\mathbf{x}}\|\}$.

déplacements le long d'une demi-droite partant de 0, des rotations classiques et des rotations hyperboliques ; chacun de ces déplacements correspond à une transformation linéaire (représentable par une matrice carrée). Voir [1] pour plus de détails sur les cônes homogènes et symétriques.

3. Considérons à présent la suite $(\mathbf{x}_k)_k \subset \mathrm{dom}(\hat{I})$ telle que

$$\mathbf{x}_k \xrightarrow[k\to+\infty]{} \mathbf{x}^* \in \mathrm{bd}\, \mathcal{Q}.$$

Par continuité de $\hat{I}$ sur son domaine, nous aurons que

$$\hat{I}(\mathbf{x}_k) \xrightarrow[k\to+\infty]{} \hat{I}(\mathbf{x}^*) = +\infty$$

puisque $\det(\mathbf{x}^*) = (x_0^*)^2 - \|\bar{\mathbf{x}}^*\|^2 = 0$.

Si nous remplaçons les inégalités du cône du second ordre $\mathbf{x}_i \succcurlyeq_{\mathcal{Q}} \mathbf{0}$ par $\mathbf{x}_i \succ_{\mathcal{Q}} \mathbf{0}$ et ajoutons le terme barrière logarithmique $-\frac{\sigma\mu}{2}\sum_{i=1}^r \ln\det(\mathbf{x}_i)$ à la fonction objectif de (P), nous obtenons

$$(P_\mu)\begin{cases} \min & \sum_{i=1}^r \mathbf{c}_i^T\mathbf{x}_i - \frac{\sigma\mu}{2}\sum_{i=1}^r \ln\det(\mathbf{x}_i), \\ \text{s.c.} & \sum_{i=1}^r A_i\mathbf{x}_i = b, \\ & \mathbf{x}_i \succ_{\mathcal{Q}} \mathbf{0},\ i = 1,\ldots,r, \end{cases} \tag{5.1}$$

où $\sigma \in [0,\, 1]$ est un paramètre de centrage et $\mu = \frac{\langle \mathbf{x}, \mathbf{z}\rangle}{r} = \frac{\sum_{i=1}^r \langle \mathbf{x}_i, \mathbf{z}_i\rangle}{r}$ est appelé le paramètre barrière. En écrivant les conditions de KKT pour le problème (P_μ), nous arrivons à :

$$\begin{array}{l} \sum_{i=1}^r A_i\mathbf{x}_i = b \\ \mathbf{c}_i - A_i^T y - \sigma\mu\mathbf{x}_i^{-1} = \mathbf{0}, \quad i = 1,\ldots,r \\ \mathbf{x}_i \succ_{\mathcal{Q}} \mathbf{0}, \quad i = 1,\ldots,r. \end{array} \tag{5.2}$$

Posons $\mathbf{z}_i = \mathbf{c}_i - A_i^T y$. Par conséquent, toute solution de (P_μ) satisfait le système suivant correspondant à une version perturbée du système d'optimalité (4.6) :

$$\begin{array}{l} \sum_{i=1}^r A_i\mathbf{x}_i = b, \\ A_i^T y + \mathbf{z}_i = \mathbf{c}_i, \quad i = 1,\ldots,r \\ \mathbf{x}_i \circ \mathbf{z}_i = \sigma\mu\mathbf{e}, \quad i = 1,\ldots,r \\ \mathbf{x}_i,\, \mathbf{z}_i \succ_{\mathcal{Q}} \mathbf{0}, \quad i = 1,\ldots,r. \end{array} \tag{5.3}$$

De façon similaire, en remplaçant $\mathbf{z}_i \succcurlyeq_{\mathcal{Q}} \mathbf{0}$ par $\mathbf{z}_i \succ_{\mathcal{Q}} \mathbf{0}$ dans le problème (D) et en ajoutant la barrière logarithmique $\frac{\sigma\mu}{2}\sum_{i=1}^r \ln\det(\mathbf{z}_i)$ à la fonction objectif duale nous obtenons

$$(D_\mu)\begin{cases} \max & b^T y + \frac{\sigma\mu}{2}\sum_{i=1}^r \ln\det(\mathbf{z}_i) \\ \text{s.c.} & A_i^T y + \mathbf{z}_i = \mathbf{c}_i, \quad i = 1,\ldots,r, \\ & \mathbf{z}_i \succ_{\mathcal{Q}} \mathbf{0}, \quad i = 1,\ldots,r. \end{cases} \tag{5.4}$$

En définissant les $\mathbf{x}_i$, pour $i = 1,\ldots,r$ comme étant les multiplicateurs de Lagrange associés aux r contraintes d'égalité de (D_μ), et en écrivant le système de KKT correspondant, il est clair que toute solution de (D_μ) satisfait aussi les équations du système (5.3).

5.3 Trajectoire centrale et direction de Newton

Il est possible de montrer (voir [11]) que, pour tout $\mu > 0$, les problèmes (P_μ) et (D_μ) possèdent une solution unique menant à une solution unique pour (5.3). Nous pouvons alors définir la notion de *trajectoire centrale primale-duale* ou plus simplement *trajectoire centrale* :

Déflnition 5.3.1 *L'ensemble des points* $(\mathrm{x}, y, \mathrm{z})$ *satisfaisant le système (5.3) pour* $\mu > 0$ *est appelé la trajectoire centrale associée au problème SOCP (1.3).*

Le principe général des méthodes du type primal-dual se basant sur des suivis de chemins est le suivant : nous partons d'un point $(\mathrm{x}^0, y^0, \mathrm{z}^0)$ proche de (ou sur) la trajectoire centrale et nous fixons un $\mu > 0$. Nous appliquons tout d'abord la méthode de Newton au système (5.3) avec ce μ fixé afin d'obtenir une direction $(\Delta\mathrm{x}, \Delta y, \Delta\mathrm{z})$ qui réduit le saut de dualité et nous nous déplaçons dans cette direction en prenant garde à ce que le nouveau point soit strictement admissible. Ensuite nous réduisons le paramètre μ d'un facteur constant et nous appliquons à nouveau la méthode de Newton au même système mais avec le μ mis à jour. Avec des choix convenables pour le point initial, la longueur de pas et le facteur de réduction pour μ, nous pourrons démontrer la convergence polynomiale des algorithmes qui se basent sur cette stratégie.
Illustrons à présent l'utilisation de la méthode de Newton. Soit $\mu > 0$. Considérons la fonction

$$\begin{aligned} F : \quad & \operatorname{int} \mathcal{Q}_{n_1} \times \ldots \times \operatorname{int} \mathcal{Q}_{n_r} \times I\!R^m \times \operatorname{int} \mathcal{Q}_{n_1} \times \ldots \times \operatorname{int} \mathcal{Q}_{n_r} \\ & \to I\!R^m \times I\!R^{n_1} \times \ldots \times I\!R^{n_r} \times I\!R^{n_1} \times \ldots \times I\!R^{n_r} \end{aligned}$$

définie par

$$(\mathrm{x}_1, \ldots, \mathrm{x}_r, y, \mathrm{z}_1, \ldots, \mathrm{z}_r) \rightsquigarrow \begin{cases} \sum_{i=1}^r A_i \mathrm{x}_i - b \\ A_i^T y + \mathrm{z}_i - \mathrm{c}_i, \quad i = 1, \ldots, r, \\ Arw(\mathrm{x}_i) Arw(\mathrm{z}_i)\mathrm{e} - \sigma\mu\mathrm{e}, \quad i = 1, \ldots, r. \end{cases}$$

L'application de la méthode de Newton à la fonction F permettra la résolution de l'équation $F(\mathrm{x}_1, \ldots, \mathrm{x}_r, y, \mathrm{z}_1, \ldots, \mathrm{z}_r) = \mathbf{0}$ et donc du système (5.3). Pour cela, il faudra résoudre le système

$$\nabla F.d_N = \begin{pmatrix} A_1 & \cdots & A_r & 0 & 0 & \cdots & 0 \\ 0 & & & A_1^T & I & & \\ & \ddots & & \vdots & & \ddots & \\ & & 0 & A_r^T & & & I \\ Arw(\mathrm{z}_1) & & & 0 & Arw(\mathrm{x}_1) & & \\ & \ddots & & \vdots & & \ddots & \\ & & Arw(\mathrm{z}_r) & 0 & & & Arw(\mathrm{x}_r) \end{pmatrix} .d_N = -F \quad (5.5)$$

où ∇F est le jacobien de F et $d_N = (\Delta\mathrm{x}_1; \ldots; \Delta\mathrm{x}_r; \Delta y; \Delta\mathrm{z}_1; \ldots; \Delta\mathrm{z}_r) = (\Delta\mathrm{x}; \Delta y; \Delta\mathrm{z})$ est la direction de Newton en $(\mathrm{x}, y, \mathrm{z})$ par rapport à $\mu > 0$. La

direction de Newton obtenue de cette manière est connue sous le nom de direction $\mathbf{xz} + \mathbf{zx}$. En réalité, au lieu déterminer directement la direction d_N en résolvant (5.5), nous allons plutôt rechercher une version mise à échelle de d_N de sorte que les matrices $Arw(\mathbf{x}_i)$ et $Arw(\mathbf{z}_i)$ correspondantes commutent ; nous pourrons alors nous assurer que cette direction est bien définie et unique.
Les vecteurs résultant du déplacement dans la direction d_N seront $\mathbf{x} + \alpha\Delta\mathbf{x}$, $y + \alpha\Delta y$ et $\mathbf{z} + \alpha\Delta\mathbf{z}$, pour une certaine longueur de pas $\alpha \in [0, 1]$.
La difficulté principale est de montrer qu'un déplacement suffisament grand à chaque itération peut être effectué dans la direction trouvée afin d'obtenir une convergence rapide. Si par exemple, à la première itération nous nous déplaçons presque aussi loin que possible dans la direction de Newton (c'est-à-dire, jusqu'à arriver près du bord de la région admissible), il se pourrait que lors des itérations ultérieures, nous ne puissions plus nous déplacer de manière efficace. Pour éviter ce problème, nous allons définir des voisinages autour de la trajectoire centrale. Ces voisinages nous permettront d'atteindre deux objectifs majeurs : premièrement, garantir un déplacement suffisament grand dans la direction de Newton sans quitter le voisinage et, deuxièment, garantir avec ce déplacement une réduction suffisament importante du saut de dualité.

Soit $\mathbf{w}_i = Q_{\mathbf{x}_i^{1/2}}\mathbf{z}_i$ et $\mathbf{w} = (\mathbf{w}_1; \ldots; \mathbf{w}_r)$. Considérons les mesures de centralité suivantes :

$$
\begin{aligned}
d_F(\mathbf{x}, \mathbf{z}) &:= \|Q_{\mathbf{x}^{1/2}}\mathbf{z} - \mu\mathbf{e}\|_F = \sqrt{\sum_{i=1}^{r}(\lambda_1(\mathbf{w}_i) - \mu)^2 + (\lambda_2(\mathbf{w}_i) - \mu)^2}, \\
d_2(\mathbf{x}, \mathbf{z}) &:= \|Q_{\mathbf{x}^{1/2}}\mathbf{z} - \mu\mathbf{e}\|_2 = \max_{i=1,\ldots,r} \max\{|\lambda_1(\mathbf{w}_i) - \mu|, |\lambda_2(\mathbf{w}_i) - \mu|\}, \\
d_{-\infty}(\mathbf{x}, \mathbf{z}) &:= \mu - \min_{i=1,\ldots,r} \min\{\lambda_1(\mathbf{w}_i), \lambda_2(\mathbf{w}_i)\}.
\end{aligned}
$$

Pour justifier ces définitions rappelons qu'en programmation linéaire les $d_\bullet(\mathbf{x}, \mathbf{z})$ sont fonctions des $x_i z_i$, et des $\lambda_i(XZ)$, (valeurs propres de XZ) en programmation semi-définie. Mais puisque les matrices XZ et $X^{1/2}ZX^{1/2}$ sont semblables, $\lambda_i(XZ) = \lambda_i(X^{1/2}ZX^{1/2})$, $\forall i$. Par ailleurs, comme nous l'avons montré à la section 3.2.4, $Q_{\mathbf{x}}$ est l'analogue de l'opérateur qui envoie toute matrice Y sur XYX dans l'algèbre des matrices symétriques. Par conséquent, $Q_{\mathbf{x}^{1/2}}\mathbf{z}$ est l'analogue de la matrice $X^{1/2}YX^{1/2}$.
En outre, la partie 2 du théorème (3.3.3) nous assure que les vecteurs $Q_{\mathbf{x}^{1/2}}\mathbf{z}$ et $Q_{\mathbf{z}^{1/2}}\mathbf{x}$ ont le même spectre, ce qui implique la symétrie des mesures $d_\bullet(\cdot, \cdot)$ par rapport à $\mathbf{x}$ et $\mathbf{z}$.
Dorénavant, nous désignerons par $\mathcal{F}^0(P)$ (resp., $\mathcal{F}^0(D)$), l'ensemble des points strictement admissibles pour le primal (resp., le dual), i.e.

$$\mathcal{F}^0(P) = \{\mathbf{x} \in I\!R^n \mid A\mathbf{x} = b,\ \mathbf{x} \succ_Q \mathbf{0}\},$$

$$\mathcal{F}^0(D) = \{(y, \mathbf{z}) \in I\!R^m \times I\!R^n \mid A^T y + \mathbf{z} = \mathbf{c},\ \mathbf{z} \succ_Q \mathbf{0}\}.$$

Ensuite, nous pouvons définir des voisinages autour de la trajectoire centrale par rapport à chacune des mesures de centralité définies. Soit $\gamma \in (0, 1)$.

$$\mathcal{N}_F(\gamma) := \{(\mathbf{x}; y; \mathbf{z}) \in \mathcal{F}^0(P) \times \mathcal{F}^0(D) \,|\, d_F(\mathbf{x}, \mathbf{z}) \leq \gamma\mu\},$$

$$\mathcal{N}_2(\gamma) := \{(\mathbf{x}; y; \mathbf{z}) \in \mathcal{F}^0(P) \times \mathcal{F}^0(D) \,|\, d_2(\mathbf{x}, \mathbf{z}) \leq \gamma\mu\},$$

$$\mathcal{N}_{-\infty}(\gamma) := \{(\mathbf{x}; y; \mathbf{z}) \in \mathcal{F}^0(P) \times \mathcal{F}^0(D) \,|\, d_{-\infty}(\mathbf{x}, \mathbf{z}) \leq \gamma\mu\}.$$

Remarquons que, puisque $d_F(\cdot, \cdot) \geq d_2(\cdot, \cdot) \geq d_{-\infty}(\cdot, \cdot)$, ces trois voisinages sont imbriqués les uns dans les autres de la façon suivante :

$$\mathcal{N}_F(\gamma) \subseteq \mathcal{N}_2(\gamma) \subseteq \mathcal{N}_{-\infty}(\gamma).$$

Donc, intuitivement, il serait préférable de travailler avec $\mathcal{N}_{-\infty}(\gamma)$ étant donné qu'il s'agit du voisinage offrant le plus d'espace pour se déplacer dans la direction de Newton. Cependant, il s'avère que pour un tel voisinage le temps de convergence vers la solution est plus important que pour les voisinages plus petits. En fait, les meilleurs résultats théoriques sont obtenus pour le plus petit voisinage, $\mathcal{N}_F(\gamma)$.

5.4 Changements d'échelle

Dans cette section nous allons voir qu'il sera possible d'effectuer des changements d'échelle sur les variables primales et duales afin d'aboutir à un formulation équivalente du système de Newton vu précédemment qui soit telle que nous retrouvions la propriété de commutativité entre les matrices $Arw(\mathbf{x})$ et $Arw(\mathbf{z})$. Soit $\mathbf{p} \succ_{\mathcal{Q}} \mathbf{0}$. A partir de $\mathbf{p}$, définissons

$$\underset{\sim}{\mathbf{u}} = Q_{\mathbf{p}}\mathbf{u} \text{ et } \underset{\sim}{\mathbf{u}} = Q_{\mathbf{p}^{-1}}\mathbf{u}.$$

Ces définitions sont valables aussi bien pour un vecteur à un seul bloc que pour un vecteur à blocs multiples. Notons que puisque $Q_{\mathbf{p}}^{-1} = Q_{\mathbf{p}^{-1}}$, les opérateurs, $\tilde{\cdot}$ et $\underset{\sim}{\cdot}$ sont inverses l'un de l'autre.
Effectuons ensuite le changement de variables : $\mathbf{x} \to \tilde{\mathbf{x}}$ qui, en vertu du théorème (3.3.2), préserve le cône $\mathcal{Q}$ invariant. Avec ce changement de variables la paire primale-duale (1.3) devient :

$$\begin{array}{llll} \widetilde{P}: & & \underset{\sim}{D}: & \\ \min & \underset{\sim}{\mathbf{c}}_1^T\tilde{\mathbf{x}}_1 + \ldots + \underset{\sim}{\mathbf{c}}_r^T\tilde{\mathbf{x}}_r & \max & b^T y \\ \text{s.c.} & \underset{\sim}{A}_1\tilde{\mathbf{x}}_1 + \ldots + \underset{\sim}{A}_r\tilde{\mathbf{x}}_r = b & \text{s.c.} & \underset{\sim}{A}_i^T y + \underset{\sim}{\mathbf{z}}_i = \underset{\sim}{\mathbf{c}}_i, \quad i = 1, \ldots, r \\ & \tilde{\mathbf{x}}_i \succcurlyeq_{\mathcal{Q}} \mathbf{0}, \quad i = 1, \ldots, r & & \underset{\sim}{\mathbf{z}}_i \succcurlyeq_{\mathcal{Q}} \mathbf{0}, \quad i = 1, \ldots, r. \end{array} \tag{5.6}$$

où

$$\mathbf{z} \to \underset{\sim}{\mathbf{z}}, \ \mathbf{c} \to \underset{\sim}{\mathbf{c}}, \ A_i \to \underset{\sim}{A}_i = A_i Q_{\mathbf{p}_i^{-1}}$$

et donc $A \to \underset{\sim}{A} = AQ_{\mathbf{p}^{-1}}$.

Lemme 5.4.1 *Pour un* $\mathbf{p} \succ_{\mathcal{Q}_n} \mathbf{0}$ *donné et sous les changements d'échelle donnés nous avons :*

1. $\mathbf{x}^T\mathbf{z} = \widetilde{\mathbf{x}}^T\underset{\sim}{\mathbf{z}}, \quad \forall \mathbf{x}, \mathbf{z} \in \mathbb{R}^n$.
2. *Pout tout vecteur* $\mathbf{u} \in \mathbb{R}^n$, $A\mathbf{u} = b$ *si et seulement si* $\underset{\sim}{A}\widetilde{\mathbf{u}} = b$.
3. $\forall \mathbf{x}, \mathbf{z} \succ_{\mathcal{Q}} \mathbf{0}$, $d_\bullet(\mathbf{x}, \mathbf{z}) = d_\bullet(\widetilde{\mathbf{x}}, \underset{\sim}{\mathbf{z}})$ *pour les mesures* F, 2 *et* $-\infty$.
4. *La trajectoire centrale et les voisinages* $\mathcal{N}_F(\cdot)$, $\mathcal{N}_2(\cdot)$ *et* $\mathcal{N}_{-\infty}(\cdot)$ *restent invariants.*

Preuve : Les parties 1 et 2 suivent directement du fait que $Q_{\mathbf{p}}^{-1} = Q_{\mathbf{p}^{-1}}$ grâce au théorème (3.3.1). Nous prouvons alors les parties 3 et 4. Posons, $\widetilde{\mathbf{w}}_i = Q_{\widetilde{\mathbf{x}}_i^{1/2}}\underset{\sim}{\mathbf{z}}_i$. Par définition, chacune des mesures $d_\bullet(\mathbf{x}, \mathbf{z})$ dépend uniquement du spectre de $\mathbf{w}_i$ et chacune des mesures $d_\bullet(\widetilde{\mathbf{x}}, \underset{\sim}{\mathbf{z}})$ dépend uniquement du spectre de $\widetilde{\mathbf{w}}_i$. Or, la première partie du théorème (3.3.3) nous assure que les vecteurs $Q_{\mathbf{x}^{1/2}}\mathbf{z}$ et $Q_{\widetilde{\mathbf{x}}^{1/2}}\underset{\sim}{\mathbf{z}}$ ont le même spectre, d'où $d_\bullet(\mathbf{x}, \mathbf{z}) = d_\bullet(\widetilde{\mathbf{x}}, \underset{\sim}{\mathbf{z}})$ pour chacune des mesures. Pour le point 4, notons que l'invariance des voisinages $\mathcal{N}_\bullet(\cdot)$ est une conséquence directe de 1, 2 et 3 ainsi que du théorème (3.3.2). Pour démontrer l'invariance de la trajectoire centrale, il nous suffit de montrer que

$$\mathbf{x} \circ \mathbf{z} = \sigma\mu\mathbf{e} \Leftrightarrow \widetilde{\mathbf{x}} \circ \underset{\sim}{\mathbf{z}} = \sigma\mu\mathbf{e}, \quad \forall \mathbf{x}, \mathbf{z} \succ_{\mathcal{Q}} \mathbf{0}.$$

Si l'implication vers la droite est démontrée, l'autre suivra immédiatement en effectuant le changement d'échelle inverse

$$\widetilde{\mathbf{x}} \to Q_{\mathbf{p}^{-1}}\widetilde{\mathbf{x}} = \mathbf{x} \text{ et } \underset{\sim}{\mathbf{z}} \to Q_{\mathbf{p}}\underset{\sim}{\mathbf{z}} = \mathbf{z}.$$

Supposons que $\mathbf{x} \circ \mathbf{z} = \sigma\mu\mathbf{e}$. Par définition, cela a pour conséquence que $\mathbf{x}^T\mathbf{z} = \sigma\mu$ et $x_0\bar{\mathbf{z}} + z_0\bar{\mathbf{x}} = \mathbf{0}$ (avec $x_0, z_0 > 0$ puisque $\mathbf{x}, \mathbf{z} \succ_{\mathcal{Q}} \mathbf{0}$). Donc, nous aurons par le point 1, $\widetilde{\mathbf{x}}^T\underset{\sim}{\mathbf{z}} = \sigma\mu$. Ensuite, montrons que $\widetilde{x}_0(\overline{\underset{\sim}{\mathbf{z}}}) + \underset{\sim}{z}_0(\overline{\widetilde{\mathbf{x}}}) = \mathbf{0}$.

$$\begin{aligned}
\underset{\sim}{z}_0(\overline{\widetilde{\mathbf{x}}}) &= (Q_{\mathbf{p}^{-1}}\mathbf{z})_0(\overline{Q_{\mathbf{p}}\mathbf{x}}) \\
&= (2(\mathbf{p}^{-1^T}\mathbf{z})(\mathbf{p}^{-1})_0 - \det(\mathbf{p}^{-1})z_0)(2(\mathbf{p}^T\mathbf{x})\bar{\mathbf{p}} + \det(\mathbf{p})\bar{\mathbf{x}}) \\
&= (\tfrac{2}{\det^2(\mathbf{p})}(p_0z_0 - \bar{\mathbf{p}}^T\bar{\mathbf{z}})p_0 - \tfrac{1}{\det(\mathbf{p})}z_0)(2(\mathbf{p}^T\mathbf{x})\bar{\mathbf{p}} - \tfrac{x_0}{z_0}\det(\mathbf{p})\bar{\mathbf{z}}) \\
&= (2(p_0z_0 - \bar{\mathbf{p}}^T\bar{\mathbf{z}})p_0 - \det(\mathbf{p})z_0)(\tfrac{2}{\det^2(\mathbf{p})}(\mathbf{p}^T\mathbf{x})\bar{\mathbf{p}} - \tfrac{x_0}{z_0}\tfrac{1}{\det(\mathbf{p})}\bar{\mathbf{z}}) \\
&= (2\tfrac{x_0}{z_0}(p_0z_0 - \bar{\mathbf{p}}^T\bar{\mathbf{z}})p_0 - \det(\mathbf{p})x_0)(\tfrac{2}{\det^2(\mathbf{p})}(p_0z_0 - \bar{\mathbf{p}}^T\bar{\mathbf{z}})\bar{\mathbf{p}} - \tfrac{1}{\det(\mathbf{p})}\bar{\mathbf{z}}) \\
&= (2(\mathbf{p}^T\mathbf{x})p_0 - \det(\mathbf{p})x_0)(-2(\mathbf{p}^{-1^T}\mathbf{z})\overline{\mathbf{p}^{-1}} - \det(\mathbf{p}^{-1})\bar{\mathbf{z}}) \\
&= (Q_{\mathbf{p}}\mathbf{x})_0(-(\overline{Q_{\mathbf{p}^{-1}}\mathbf{z}})) \\
&= -\widetilde{x}_0(\overline{\underset{\sim}{\mathbf{z}}})
\end{aligned}$$

Par conséquent, nous obtenons bien : $\widetilde{\mathbf{x}} \circ \underset{\sim}{\mathbf{z}} = (\widetilde{\mathbf{x}}^T\underset{\sim}{\mathbf{z}};\ \widetilde{x}_0(\overline{\underset{\sim}{\mathbf{z}}}) + \underset{\sim}{z}_0(\overline{\widetilde{\mathbf{x}}})) = \sigma\mu\mathbf{e}$.

□

Lemme 5.4.2 *Pour un* $\mathbf{p} \succ_{\mathcal{Q}} \mathbf{0}$ *fixé, le triplet de directions* $(\widetilde{\Delta \mathbf{x}}, \Delta y, \underset{\sim}{\Delta \mathbf{z}})$ *résoud le système d'équations*

$$\begin{aligned} \underset{\sim}{A}\widetilde{\Delta \mathbf{x}} &= b - \underset{\sim}{A}\widetilde{\mathbf{x}} \\ \underset{\sim}{A}^T \Delta y + \underset{\sim}{\Delta \mathbf{z}} &= \underset{\sim}{\mathbf{c}} - \underset{\sim}{A}^T y - \underset{\sim}{\mathbf{z}} \\ \widetilde{\Delta \mathbf{x}} \circ \underset{\sim}{\mathbf{z}} + \widetilde{\mathbf{x}} \circ \underset{\sim}{\Delta \mathbf{z}} &= \sigma\mu \mathbf{e} - \widetilde{\mathbf{x}} \circ \underset{\sim}{\mathbf{z}} \end{aligned} \tag{5.7}$$

si et seulement si $(\Delta \mathbf{x}, \Delta y, \Delta \mathbf{z})$ *résoud*

$$\begin{aligned} A\Delta \mathbf{x} &= b - A\mathbf{x} \\ A^T \Delta y + \Delta \mathbf{z} &= \mathbf{c} - A^T y - \mathbf{z} \\ (Q_{\mathbf{p}} \Delta \mathbf{x}) \circ (Q_{\mathbf{p}^{-1}} \mathbf{z}) + (Q_{\mathbf{p}} \mathbf{x}) \circ (Q_{\mathbf{p}^{-1}} \Delta \mathbf{z}) &= \sigma\mu \mathbf{e} - (Q_{\mathbf{p}} \mathbf{x}) \circ (Q_{\mathbf{p}^{-1}} \mathbf{z}). \end{aligned} \tag{5.8}$$

Remarquons que la direction $(\widetilde{\Delta \mathbf{x}}, \Delta y, \underset{\sim}{\Delta \mathbf{z}})$ est identique à celle obtenue par application de la méthode de Newton au système (5.3) avec les nouvelles variables. La direction $(\Delta \mathbf{x}, \Delta y, \Delta \mathbf{z})$ qui lui correspond dans (5.8) est différente de celle obtenue par résolution du système (5.5), c'est-à-dire que pour tout vecteur $\mathbf{p} \succ_{\mathcal{Q}} \mathbf{0}$ le système (5.8) nous donnera une direction particulière, différente, en général, de la direction $\mathbf{xz} + \mathbf{zx}$. Cette direction est obtenue dans le cas particulier où $\mathbf{p}$ est un multiple strictement positif de $\mathbf{e}$ (par exemple, $\mathbf{p} = \mathbf{e}$).

5.5 Directions commutatives

Nous allons nous intéresser de plus près à la classe des vecteurs $\mathbf{p} \succ_{\mathcal{Q}} \mathbf{0}$ pour lesquels les nouvelles variables $\widetilde{\mathbf{x}}$ et $\underset{\sim}{\mathbf{z}}$ commutent.

Définition 5.5.1 *L'ensemble des directions* $(\Delta \mathbf{x}, \Delta y, \Delta \mathbf{z})$ *obtenues par (5.8) avec des vecteurs* $\mathbf{p} \succ_{\mathcal{Q}} \mathbf{0}$ *tels que* $\widetilde{\mathbf{x}}$ *et* $\underset{\sim}{\mathbf{z}}$ *commutent est appelé la classe commutative des directions ; une direction de cette classe est appelée une direction commutative.*

Remarque : L'ensemble des vecteurs $\mathbf{p} \succ_{\mathcal{Q}} \mathbf{0}$ tels que $\widetilde{\mathbf{x}}$ et $\underset{\sim}{\mathbf{z}}$ commutent est noté $\mathcal{C}(\mathbf{x}, \mathbf{z})$.
Il est clair que puisque $\mathbf{x}$ et $\mathbf{z}$ ne commutent pas en général, la direction résultant du choix $\mathbf{p} = \mathbf{e}$ (i.e., $\widetilde{\mathbf{x}} = \mathbf{x}$ et $\underset{\sim}{\mathbf{z}} = \mathbf{z}$) n'est pas commutative. Cependant, les choix suivants pour $\mathbf{p}$ engendrent tous des directions commutatives :

$$\mathbf{p} = \mathbf{z}^{1/2}, \qquad \mathbf{p} = \mathbf{x}^{1/2},$$
$$\mathbf{p} = \left[Q_{\mathbf{x}^{1/2}}(Q_{\mathbf{x}^{1/2}}\mathbf{z})^{-1/2}\right]^{-1/2}.$$

Les directions résultant des deux premiers choix sont appelées respectivement directions $\mathbf{xz}$ et $\mathbf{zx}$ par analogie à la programmation semi-définie. Le troisième

choix pour $\mathbf{p}$ résulte en la direction appelée direction de Nesterov-Todd. Notons que le théorème (3.3.2) combiné au fait que $\mathbf{x}$ et $\mathbf{z}$ appartiennent à int $\mathcal{Q}$, nous assurent que chacun des $\mathbf{p}$ définis appartient à int $\mathcal{Q}$. Vérifions à présent que ces vecteurs $\mathbf{p}$ nous donnent des $\widetilde{\mathbf{x}}$ et $\underset{\sim}{\mathbf{z}}$ qui commutent :

$\underline{\mathbf{p} = \mathbf{z}^{1/2}}$: Grâce aux parties 3 et 5 du théorème (3.3.1) nous pouvons écrire

$$\underset{\sim}{\mathbf{z}} = Q_{\mathbf{p}^{-1}}\mathbf{z} = Q_{\mathbf{z}^{-1/2}}(\mathbf{z}^{1/2})^2 = (Q_{\mathbf{z}^{-1/2}}Q_{\mathbf{z}^{1/2}})\mathbf{e} = \mathbf{e}.$$

$\underline{\mathbf{p} = \mathbf{x}^{-1/2}}$: En utilisant les mêmes résultats nous avons

$$\widetilde{\mathbf{x}} = Q_{\mathbf{p}}\mathbf{x} = Q_{\mathbf{x}^{-1/2}}(\mathbf{x}^{1/2})^2 = (Q_{\mathbf{x}^{-1/2}}Q_{\mathbf{x}^{1/2}})\mathbf{e} = \mathbf{e}.$$

Direction de Nesterov-Todd : Nous avons tout d'abord

$$\begin{aligned}
Q_{\mathbf{p}}^2\mathbf{x} &= Q_{\mathbf{p}^2}\mathbf{x} \\
&= (Q_{[Q_{\mathbf{x}^{1/2}}(Q_{\mathbf{x}^{1/2}}\mathbf{z})^{-1/2}]^{-1}})\mathbf{x} \\
&= (Q_{Q_{\mathbf{x}^{-1/2}}(Q_{\mathbf{x}^{1/2}}\mathbf{z})^{1/2}})\mathbf{x} \text{ (théorème (3.3.1), point 6)} \\
&= (Q_{\mathbf{x}^{-1/2}}Q_{(Q_{\mathbf{x}^{1/2}}\mathbf{z})^{1/2}}Q_{\mathbf{x}^{-1/2}})\mathbf{x} \text{ (théorème (3.3.1), point 8)} \\
&= (Q_{\mathbf{x}^{-1/2}}Q_{(Q_{\mathbf{x}^{1/2}}\mathbf{z})^{1/2}}Q_{\mathbf{x}^{-1/2}})Q_{\mathbf{x}^{1/2}}\mathbf{e} \text{ (théorème (3.3.1), point 3)} \\
&= (Q_{\mathbf{x}^{-1/2}}Q_{(Q_{\mathbf{x}^{1/2}}\mathbf{z})^{1/2}})\mathbf{e} \\
&= Q_{\mathbf{x}^{-1/2}}(Q_{\mathbf{x}^{1/2}}\mathbf{z}) \\
&= \mathbf{z}.
\end{aligned}$$

D'où, $\widetilde{\mathbf{x}} = Q_{\mathbf{p}}\mathbf{x} = Q_{\mathbf{p}^{-1}}Q_{\mathbf{p}^2}\mathbf{x} = Q_{\mathbf{p}^{-1}}\mathbf{z} = \underset{\sim}{\mathbf{z}}$.

Il suit que dans chacun des cas, les vecteurs $\widetilde{\mathbf{x}}$ et $\underset{\sim}{\mathbf{z}}$ commutent. Par la même occasion, nous obtenons grâce au théorème (3.2.2) la commutativité entre les matrices $Arw(\widetilde{\mathbf{x}})$ et $Arw(\underset{\sim}{\mathbf{z}})$. Nous sommes alors en mesure de montrer la proposition suivante

Proposition 5.5.1 *Soit* $(\mathbf{x}, y, \mathbf{z})$ *un point admissible pour* (P) *et* (D). *Si* $\mathbf{x} \succ_{\mathcal{Q}} \mathbf{0}$ *et* $\mathbf{z} \succ_{\mathcal{Q}} \mathbf{0}$, *alors la direction de Newton mise à échelle* $(\widetilde{\Delta\mathbf{x}}, \Delta y, \underset{\sim}{\Delta\mathbf{z}})$ *obtenue par résolution de (5.7) pour un certain* $p \in \mathcal{C}(\mathbf{x}, \mathbf{z})$ *est bien définie et unique.*

Preuve : Puisque $\mathbf{x} \succ_{\mathcal{Q}} \mathbf{0}$ et $\mathbf{z} \succ_{\mathcal{Q}} \mathbf{0}$, nous avons $\widetilde{\mathbf{x}} \succ_{\mathcal{Q}} \mathbf{0}$ et $\underset{\sim}{\mathbf{z}} \succ_{\mathcal{Q}} \mathbf{0}$. Les matrices $Arw(\widetilde{\mathbf{x}})$ et $Arw(\underset{\sim}{\mathbf{z}})$ sont alors définies positives et donc non-singulières. De plus, ces deux matrices commutent entre elles grâce au choix particulier de $\mathbf{p}$. Le résultat annoncé sera démontré si nous parvenons à montrer la non-singularité de la matrice

$$\begin{pmatrix} \underset{\sim}{A} & 0 & 0 \\ 0 & \underset{\sim}{A}^T & I \\ Arw(\underset{\sim}{\mathbf{z}}) & 0 & Arw(\widetilde{\mathbf{x}}) \end{pmatrix}.$$

Pour prouver cela, nous considérons $(u, v, w) \in I\!R^n \times I\!R^m \times I\!R^n$ tel que

$$\underset{\sim}{A}u = 0, \quad \underset{\sim}{A}^T v + w = 0, \quad Arw(\underset{\sim}{z})u + Arw(\widetilde{\mathbf{x}})w = 0$$

et nous montrons que $u = 0$, $v = 0$ et $w = 0$. Remarquons tout d'abord que

$$Arw(\widetilde{\mathbf{x}})\underset{\sim}{A}^T v = -Arw(\widetilde{\mathbf{x}})w = Arw(\underset{\sim}{z})u$$

et

$$u = Arw^{-1}(\underset{\sim}{z})Arw(\widetilde{\mathbf{x}})\underset{\sim}{A}^T v$$

de sorte que nous ayons $\underset{\sim}{A}u = \underset{\sim}{A}Arw^{-1}(\underset{\sim}{z})Arw(\widetilde{\mathbf{x}})\underset{\sim}{A}^T v = 0$ ainsi que $v^T\underset{\sim}{A}Arw^{-1}(\underset{\sim}{z})Arw(\widetilde{\mathbf{x}})\underset{\sim}{A}^T v = 0$. Puisque

$$Arw^{-1}(\underset{\sim}{z})Arw(\widetilde{\mathbf{x}}) = Arw^{-1/2}(\underset{\sim}{z})Arw^{1/2}(\widetilde{\mathbf{x}})Arw^{1/2}(\widetilde{\mathbf{x}})Arw^{-1/2}(\underset{\sim}{z}),$$

nous avons $\|Arw^{1/2}(\widetilde{\mathbf{x}})Arw^{-1/2}(\underset{\sim}{z})\underset{\sim}{A}^T v\|^2 = 0$ et donc

$$Arw^{1/2}(\widetilde{\mathbf{x}})Arw^{-1/2}(\underset{\sim}{z})\underset{\sim}{A}^T v = 0 \text{ et } \underset{\sim}{A}^T v = 0.$$

Mais alors $v = 0$ car A (et donc $\underset{\sim}{A}$) est de rang plein. Par conséquent, $u = 0$ et $w = 0$. Donc, la direction de Newton mise à échelle $(\widetilde{\Delta\mathbf{x}}, \Delta y, \underset{\sim}{\Delta z})$ est bien définie et unique.

□

5.6 Algorithmes de suivi de chemins

Cette section est consacrée au développement des algorithmes de points intérieurs du type primal-dual pour des problèmes SOCP. Ces algorithmes vont se baser sur la stratégie de suivi de chemins et vont exploiter les propriétés des voisinages que nous avons définis.

Supposons que $\mathbf{x} \in \mathcal{F}^0(P)$ et $(y, \mathbf{z}) \in \mathcal{F}^0(D)$, que nous ayons $\mathbf{p} \in \mathcal{C}(\mathbf{x}, \mathbf{z})$ et que $(\Delta\mathbf{x}, \Delta y, \Delta\mathbf{z})$ est la solution de (5.8). Nous utiliserons les notations suivantes

$$\begin{aligned}
&\mathbf{x}(\alpha) = \mathbf{x} + \alpha\Delta\mathbf{x}, \quad \mathbf{z}(\alpha) = \mathbf{z} + \alpha\Delta\mathbf{z},\\
&\widetilde{\mathbf{x}}(\alpha) = \widetilde{\mathbf{x}} + \alpha\widetilde{\Delta\mathbf{x}}, \quad \underset{\sim}{z}(\alpha) = \underset{\sim}{z} + \alpha\underset{\sim}{\Delta z},\\
&\mu(\alpha) = \mu(\mathbf{x}(\alpha), \mathbf{z}(\alpha)) = \frac{\langle \mathbf{x}(\alpha), \mathbf{z}(\alpha)\rangle}{r},\\
&\widetilde{\mathbf{w}}(\alpha) = Q_{\widetilde{\mathbf{x}}(\alpha)^{1/2}}\underset{\sim}{z}(\alpha),
\end{aligned}$$

où $\alpha \in [0, 1]$ représente la longueur de pas du déplacement dans la direction $(\Delta\mathbf{x}, \Delta y, \Delta\mathbf{z})$. Le lemme qui va suivre nous renseigne sur la manière dont évolue le saut de dualité lorsqu'à partir d'un point strictement admissible $(\mathbf{x}, y, \mathbf{z})$, nous nous déplaçons dans la direction $(\Delta\mathbf{x}, \Delta y, \Delta\mathbf{z})$. Nous utiliserons de manière définitive la notation $\langle\cdot,\cdot\rangle$ pour le produit scalaire dans $I\!R^n$.

Lemme 5.6.1 *Soient* $(\mathbf{x}, \mathbf{z}) \in \mathcal{F}^0(P) \times \mathcal{F}^0(D)$ *et la direction* $(\Delta\mathbf{x}, \Delta y, \Delta\mathbf{z})$ *obtenue avec (5.8). Alors, nous avons*

1. $\langle \widetilde{\Delta\mathbf{x}}, \underset{\sim}{\Delta\mathbf{z}} \rangle = \langle \Delta\mathbf{x}, \Delta\mathbf{z} \rangle = 0,$
2. $\mu(\alpha) = [1 - \alpha(1 - \sigma)]\mu.$

Preuve : Lorsque $(\mathbf{x}, \mathbf{z}) \in \mathcal{F}^0(P) \times \mathcal{F}^0(D)$, nous obtenons à partir du système (5.8) que $(\Delta\mathbf{x}, \Delta\mathbf{z}) \in \operatorname{Ker} A \times \operatorname{Im} A^T$. Donc, puisque $(\operatorname{Ker} A)^\perp = \operatorname{Im} A^T$, nous avons $\langle \Delta\mathbf{x}, \Delta\mathbf{z} \rangle = 0$. La première égalité dans 1 est établie grâce au point 1 du lemme (5.4.1).

Pour le point 2, nous avons

$$\begin{aligned}
\langle \mathbf{x}(\alpha), z(\alpha) \rangle &= \langle \widetilde{\mathbf{x}}(\alpha), \underset{\sim}{z}(\alpha) \rangle \\
&= \langle \widetilde{\mathbf{x}}, \underset{\sim}{z} \rangle + \alpha(\langle \widetilde{\mathbf{x}}, \underset{\sim}{\Delta z} \rangle + \langle \underset{\sim}{z}, \widetilde{\Delta\mathbf{x}} \rangle) + \alpha^2 \langle \widetilde{\Delta\mathbf{x}}, \underset{\sim}{\Delta z} \rangle \\
&= (1 - \alpha)\langle \widetilde{\mathbf{x}}, \underset{\sim}{z} \rangle + \alpha\langle \widetilde{\mathbf{x}}, \underset{\sim}{z} \rangle + \alpha\langle \widetilde{\mathbf{x}} \circ \underset{\sim}{\Delta z} + \underset{\sim}{z} \circ \widetilde{\Delta\mathbf{x}}, \mathbf{e} \rangle \\
&= (1 - \alpha)\langle \widetilde{\mathbf{x}}, \underset{\sim}{z} \rangle + \alpha\langle \widetilde{\mathbf{x}} \circ \underset{\sim}{\Delta z} + \underset{\sim}{z} \circ \widetilde{\Delta\mathbf{x}} + \widetilde{\mathbf{x}} \circ \underset{\sim}{z}, \mathbf{e} \rangle \\
&= (1 - \alpha)\langle \widetilde{\mathbf{x}}, \underset{\sim}{z} \rangle + \alpha\langle \sigma\mu\mathbf{e}, \mathbf{e} \rangle \\
&= (1 - \alpha)\langle \mathbf{x}, \mathbf{z} \rangle + \alpha\sigma\mu r \\
&= r[(1 - \alpha)\mu + \alpha\sigma\mu] \\
&= r\mu[1 - \alpha(1 - \sigma)]
\end{aligned}$$

$$\Rightarrow \mu(\alpha) = \frac{\langle \mathbf{x}(\alpha), z(\alpha) \rangle}{r} = (1 - \alpha(1 - \sigma))\mu.$$

□

En termes d'itérations, nous avons

$$\mu_{k+1} = [1 - \alpha_k(1 - \sigma_k)]\mu_k, \quad \forall k.$$

Ainsi, lorsque $\sigma_k = 1$, nous avons $\mu_{k+1} = \mu_k$, i.e., le saut de dualité ne décroît pas. Si $\sigma_k = 0$, alors $\mu_{k+1} = [1 - \alpha_k]\mu_k$, i.e., le saut de dualité décroît. Enfin, si $\alpha_k = 1$, i.e., déplacements de type Newton pure, alors $\mu_{k+1} = \sigma_k\mu_k$.

5.6.1 Lemmes techniques

Lemme 5.6.2 *Soit* $(\mathbf{x}, \mathbf{z}) \in int\,\mathcal{Q} \times int\,\mathcal{Q}$. *Alors, les inégalités suivantes sont vérifiées :*

$$\|\mathbf{x} \circ \mathbf{z} - \mu\mathbf{e}\|_F \geq \|\mathbf{w} - \mu\mathbf{e}\|_F \tag{5.9}$$

$$\|\mathbf{x} \circ \mathbf{z} - \mu\mathbf{e}\|_2 \geq \|\mathbf{w} - \mu\mathbf{e}\|_2 \tag{5.10}$$

$$\mu - \min_{\substack{i=1,\ldots,r \\ j=1,2}} \lambda^j(\mathbf{x}_i \circ \mathbf{z}_i) \geq \mu - \min_{\substack{i=1,\ldots,r \\ j=1,2}} \lambda^j(\mathbf{w}_i) \tag{5.11}$$

où les inégalités deviennent des égalités lorsque $\mathbf{x}$ *et* $\mathbf{z}$ *commutent.*

Preuve : Pour (5.9), notons que puisque les matrices $Arw(\mathbf{x})$ et $Arw(\mathbf{x}^2)$ commutent, il en est de même pour $Arw^2(\mathbf{x})$ et $Arw(\mathbf{x}^2)$. Ces deux matrices possèdent alors des vecteurs propres identiques, ce qui nous permet d'écrire :

$$Arw(\mathbf{x}^2) = P\Omega P^T \text{ et } Arw^2(\mathbf{x}) = P\Lambda^2 P^T$$

où nous avons utilisé les mêmes notations que dans la preuve du théorème (3.2.1). Il est alors clair que $\Omega - \Lambda^2 \succcurlyeq 0$ et donc que $Arw(\mathbf{x}^2) - Arw^2(\mathbf{x}) \succcurlyeq 0$. Ainsi, nous avons :

$$\begin{aligned}
\|\mathbf{x} \circ \mathbf{z} - \mu\mathbf{e}\|^2 &= \langle \mathbf{x} \circ \mathbf{z}, \mathbf{x} \circ \mathbf{z}\rangle - 2\mu\langle \mathbf{x} \circ \mathbf{z}, \mathbf{e}\rangle + \mu^2\langle \mathbf{e}, \mathbf{e}\rangle \\
&= \langle \mathbf{z}, Arw^2(\mathbf{x})\mathbf{z}\rangle - 2\mu\langle \mathbf{x}, \mathbf{z}\rangle + \mu^2\langle \mathbf{e}, \mathbf{e}\rangle \\
&= \langle \mathbf{z}, 2Arw^2(\mathbf{x})\mathbf{z}\rangle - \langle \mathbf{z}, Arw^2(\mathbf{x})\mathbf{z}\rangle - 2\mu\langle \mathbf{x}, \mathbf{z}\rangle + \mu^2\langle \mathbf{e}, \mathbf{e}\rangle \\
&\geq \langle \mathbf{z}, 2Arw^2(\mathbf{x})\mathbf{z}\rangle - \langle \mathbf{z}, Arw(\mathbf{x}^2)\mathbf{z}\rangle - 2\mu\langle \mathbf{x}, \mathbf{z}\rangle + \mu^2\langle \mathbf{e}, \mathbf{e}\rangle \\
&= \langle \mathbf{z}, (2Arw^2(\mathbf{x}) - Arw(\mathbf{x}^2))\mathbf{z}\rangle - 2\mu\langle Q_{\mathbf{x}^{1/2}}\mathbf{e}, \mathbf{z}\rangle + \mu^2\langle \mathbf{e}, \mathbf{e}\rangle \\
&= \langle \mathbf{z}, Q_{\mathbf{x}}\mathbf{z}\rangle - 2\mu\langle Q_{\mathbf{x}^{1/2}}\mathbf{e}, \mathbf{z}\rangle + \mu^2\langle \mathbf{e}, \mathbf{e}\rangle \\
&= \langle Q_{\mathbf{x}^{1/2}}\mathbf{z}, Q_{\mathbf{x}^{1/2}}\mathbf{z}\rangle - 2\mu\langle \mathbf{e}, Q_{\mathbf{x}^{1/2}}\mathbf{z}\rangle + \mu^2\langle \mathbf{e}, \mathbf{e}\rangle \\
&= \|\mathbf{w} - \mu\mathbf{e}\|^2.
\end{aligned}$$

Par conséquent, puisque $\|\cdot\|_F = \sqrt{2}\|\cdot\|$, l'inégalité (5.9) est démontrée.

Pour (5.10) et (5.11), il suffit de montrer que

$$\lambda_{\min}(\mathbf{x} \circ \mathbf{z}) \leq \lambda_{\min}(\mathbf{w}) \text{ et } \lambda_{\max}(\mathbf{x} \circ \mathbf{z}) \geq \lambda_{\max}(\mathbf{w})$$

(avec $\lambda_{\max}(\cdot_i) \equiv \max\limits_{\substack{i=1,\ldots,r \\ j=1,2}} \lambda^j(\cdot_i)$ et $\lambda_{\min}(\cdot_i) \equiv \min\limits_{\substack{i=1,\ldots,r \\ j=1,2}} \lambda^j(\cdot_i)$) car ainsi nous aurons d'une part,

$$\begin{aligned}
\|\mathbf{x} \circ \mathbf{z} - \mu\mathbf{e}\|_2 &= \max_{\substack{i=1,\ldots,r \\ j=1,2}} |\lambda^j(\mathbf{x}_i \circ \mathbf{z}_i) - \mu| \geq \max_{\substack{i=1,\ldots,r \\ j=1,2}} (\lambda^j(\mathbf{x}_i \circ \mathbf{z}_i) - \mu) \\
&= \lambda_{\max}(\mathbf{x} \circ \mathbf{z}) - \mu \geq \lambda_{\max}(\mathbf{w}) - \mu \\
&= \max_{\substack{i=1,\ldots,r \\ j=1,2}} (\lambda^j(\mathbf{w}_i) - \mu), \ (\equiv A)
\end{aligned}$$

et d'autre part,

$$\begin{aligned}
\|\mathbf{x} \circ \mathbf{z} - \mu\mathbf{e}\|_2 &= \max_{\substack{i=1,\ldots,r \\ j=1,2}} |\lambda^j(\mathbf{x}_i \circ \mathbf{z}_i) - \mu| \geq \max_{\substack{i=1,\ldots,r \\ j=1,2}} (\mu - \lambda^j(\mathbf{x}_i \circ \mathbf{z}_i)) \\
&= \mu + \max_{\substack{i=1,\ldots,r \\ j=1,2}} (-\lambda^j(\mathbf{x}_i \circ \mathbf{z}_i)) = \mu - \lambda_{\min}(\mathbf{x} \circ \mathbf{z}) \\
&\geq \mu - \lambda_{\min}(\mathbf{w}) = \mu + \max_{\substack{i=1,\ldots,r \\ j=1,2}} (-\lambda^j(\mathbf{w}_i)) \\
&= \max_{\substack{i=1,\ldots,r \\ j=1,2}} (\mu - \lambda^j(\mathbf{w}_i)). \ (\equiv B)
\end{aligned}$$

D'où, $\|\mathbf{x} \circ \mathbf{z} - \mu\mathbf{e}\|_2$ sera plus grand ou égal que le maximum entre A et B, qui équivaut à $\|\mathbf{w} - \mu\mathbf{e}\|_2$. L'inégalité (5.11) suivra de façon évidente.

Montrons donc $\lambda_{\min}(\mathbf{x} \circ \mathbf{z}) \leq \lambda_{\min}(\mathbf{w})$ et $\lambda_{\max}(\mathbf{x} \circ \mathbf{z}) \geq \lambda_{\max}(\mathbf{w})$.
En utilisant la propriété 10 du théorème (3.3.1) avec $\mathbf{x} \hookleftarrow \mathbf{x}^{1/2}$ nous obtenons

$$Q_{\mathbf{x}^{1/2}, \mathbf{x}^{-1/2}}\mathbf{w} = Q_{\mathbf{x}^{1/2}, \mathbf{x}^{-1/2}} Q_{\mathbf{x}^{1/2}}\mathbf{z} = Arw(\mathbf{x})\mathbf{z} = \mathbf{x} \circ \mathbf{z}.$$

De plus,

$$\begin{aligned}
\mathrm{tr}(Q_{\mathbf{x}^{1/2}, \mathbf{x}^{-1/2}}\mathbf{u}) &= \mathrm{tr}(2Arw(\mathbf{x}^{1/2})Arw(\mathbf{x}^{-1/2})\mathbf{u} - Arw(\mathbf{e})\mathbf{u}) \\
&= 2\mathrm{tr}(Arw(\mathbf{x}^{1/2})Arw(\mathbf{x}^{-1/2})\mathbf{u}) - \mathrm{tr}(Arw(\mathbf{e})\mathbf{u}) \\
&= 2\mathrm{tr}(\mathbf{x}^{1/2} \circ (\mathbf{x}^{-1/2} \circ \mathbf{u})) - \mathrm{tr}(Arw(\mathbf{e})\mathbf{u}) \\
&= 2(2\langle \mathbf{x}^{1/2}, (\mathbf{x}^{-1/2} \circ \mathbf{u})\rangle) - \mathrm{tr}(\mathbf{u}) \\
&= 2(2\langle \mathbf{x}^{1/2}, Arw(\mathbf{x}^{-1/2})\mathbf{u}\rangle) - \mathrm{tr}(\mathbf{u}) \\
&= 2(2\langle Arw(\mathbf{x}^{-1/2})\mathbf{x}^{1/2}, \mathbf{u}\rangle) - \mathrm{tr}(\mathbf{u}) \\
&= 2(2\langle \mathbf{e}, \mathbf{u}\rangle) - \mathrm{tr}(\mathbf{u}) \\
&= 2\mathrm{tr}(\mathbf{e} \circ \mathbf{u}) - \mathrm{tr}(\mathbf{u}) = \mathrm{tr}(\mathbf{u}).
\end{aligned}$$

Cela nous permet d'obtenir

$$\begin{aligned}
\lambda_{\min}(\mathbf{x} \circ \mathbf{z}) &= \lambda_{\min}(Arw(\mathbf{x} \circ \mathbf{z})) = \min_{\mathbf{u} \neq 0} \frac{\langle \mathbf{u}, Arw(\mathbf{x} \circ \mathbf{z})\mathbf{u}\rangle}{\langle \mathbf{u}, \mathbf{u}\rangle} \\
&= \min_{\mathbf{u} \neq 0} \frac{\langle \mathbf{u}, (\mathbf{x} \circ \mathbf{z}) \circ \mathbf{u}\rangle}{\langle \mathbf{u}, \mathbf{u}\rangle} = \min_{\mathbf{u} \neq 0} \frac{\langle \mathbf{u}, \mathbf{u} \circ (\mathbf{x} \circ \mathbf{z})\rangle}{\langle \mathbf{u}, \mathbf{u}\rangle} \\
&= \min_{\mathbf{u} \neq 0} \frac{\langle \mathbf{u}, Arw(\mathbf{u})(\mathbf{x} \circ \mathbf{z})\rangle}{\langle \mathbf{u}, \mathbf{u}\rangle} = \min_{\mathbf{u} \neq 0} \frac{\langle \mathbf{x} \circ \mathbf{z}, Arw(\mathbf{u})\mathbf{u}\rangle}{\langle \mathbf{u}, \mathbf{u}\rangle} \\
&= \min_{\mathbf{u} \neq 0} \frac{\langle \mathbf{x} \circ \mathbf{z}, \mathbf{u}^2\rangle}{\langle \mathbf{u}, \mathbf{u}\rangle} = \min_{\langle \mathbf{u}, \mathbf{u}\rangle = 1} \langle \mathbf{u}^2, \mathbf{x} \circ \mathbf{z}\rangle \\
&= \min_{\mathrm{tr}(\mathbf{u}^2)=2} \langle \mathbf{u}^2, Q_{\mathbf{x}^{1/2}, \mathbf{x}^{-1/2}}\mathbf{w}\rangle = \min_{\mathrm{tr}(\mathbf{u}^2)=2} \langle Q_{\mathbf{x}^{1/2}, \mathbf{x}^{-1/2}}\mathbf{u}^2, \mathbf{w}\rangle \\
&\leq \min_{\substack{Q_{\mathbf{x}^{1/2}, \mathbf{x}^{-1/2}}\mathbf{u}^2 = \mathbf{t}^2 \\ \mathrm{tr}(\mathbf{t}^2)=2}} \langle \mathbf{t}^2, \mathbf{w}\rangle = \min_{\mathrm{tr}(\mathbf{t}^2)=2} \langle \mathbf{t}^2, \mathbf{w}\rangle \\
&= \min_{\langle \mathbf{t}, \mathbf{t}\rangle = 1} \langle Arw(\mathbf{t})\mathbf{t}, \mathbf{w}\rangle = \min_{\langle \mathbf{t}, \mathbf{t}\rangle = 1} \langle \mathbf{t}, Arw(\mathbf{t})\mathbf{w}\rangle \\
&= \min_{\langle \mathbf{t}, \mathbf{t}\rangle = 1} \langle \mathbf{t}, \mathbf{w} \circ \mathbf{t}\rangle = \min_{\langle \mathbf{t}, \mathbf{t}\rangle = 1} \langle \mathbf{t}, Arw(\mathbf{w})\mathbf{t}\rangle \\
&= \lambda_{\min}(\mathbf{w})
\end{aligned}$$

De façon similaire

$$\begin{aligned}
\lambda_{\max}(\mathbf{x} \circ \mathbf{z}) &= \max_{\mathrm{tr}(\mathbf{u}^2)=2} \langle Q_{\mathbf{x}^{1/2}, \mathbf{x}^{-1/2}}\mathbf{u}^2, \mathbf{w}\rangle \geq \max_{\substack{Q_{\mathbf{x}^{1/2}, \mathbf{x}^{-1/2}}\mathbf{u}^2 = \mathbf{t}^2 \\ \mathrm{tr}(\mathbf{t}^2)=2}} \langle \mathbf{t}^2, \mathbf{w}\rangle \\
&= \max_{\mathrm{tr}(\mathbf{t}^2)=2} \langle \mathbf{t}^2, \mathbf{w}\rangle = \lambda_{\max}(\mathbf{w}).
\end{aligned}$$

Pour les égalités, notons que si $\mathbf{x}$ et $\mathbf{z}$ commutent il en est de même pour $\mathbf{x}^{1/2}$ et $\mathbf{z}$ et, de manière équivalente, pour les matrices $Arw(\mathbf{x}^{1/2})$ et $Arw(\mathbf{z})$.

Dès lors,

$$\begin{aligned}\mathbf{w} &= Q_{\mathbf{x}^{1/2}}\mathbf{z} = (2Arw^2(\mathbf{x}^{1/2}) - Arw(\mathbf{x}))\mathbf{z} \\ &= 2Arw^2(\mathbf{x}^{1/2})Arw(\mathbf{z})\mathbf{e} - \mathbf{x}\circ\mathbf{z} \\ &= 2Arw(\mathbf{z})Arw^2(\mathbf{x}^{1/2})\mathbf{e} - \mathbf{x}\circ\mathbf{z} \\ &= 2Arw(\mathbf{z})\mathbf{x} - \mathbf{x}\circ\mathbf{z} = \mathbf{x}\circ\mathbf{z}.\end{aligned}$$

□

Lemme 5.6.3 *Soient* $\mathbf{x}, \mathbf{y} \in I\!R^n$, *tels que* $\mathbf{x} \equiv (\mathbf{x}_1, \ldots, \mathbf{x}_r)$ *et* $\mathbf{y} \equiv (\mathbf{y}_1, \ldots, \mathbf{y}_r)$ *avec* $\mathbf{x}_i, \mathbf{y}_i \in I\!R^{n_i}$. *Alors nous avons les bornes suivantes pour les valeurs propres de* $\mathbf{x}+\mathbf{y}$:

1. $\lambda_{\min}(\mathbf{x}+\mathbf{y}) \geq \lambda_{\min}(\mathbf{x}) - \|\mathbf{y}\|_F$,
2. $\lambda_{\max}(\mathbf{x}+\mathbf{y}) \leq \lambda_{\max}(\mathbf{x}) + \|\mathbf{y}\|_F$.

Preuve : Pour 1, nous avons

$$\begin{aligned}\lambda_{\min}(\mathbf{x}+\mathbf{y}) &= \min_{\mathbf{u}\neq 0} \frac{\langle \mathbf{u}, (\mathbf{x}+\mathbf{y})\circ\mathbf{u}\rangle}{\langle \mathbf{u}, \mathbf{u}\rangle} \\ &= \min_{\mathbf{u}\neq 0} \frac{\langle \mathbf{u}, \mathbf{x}\circ\mathbf{u}\rangle + \langle \mathbf{u}, \mathbf{y}\circ\mathbf{u}\rangle}{\langle \mathbf{u}, \mathbf{u}\rangle} \\ &\geq \min_{\mathbf{u}\neq 0} \frac{\langle \mathbf{u}, \mathbf{x}\circ\mathbf{u}\rangle}{\langle \mathbf{u}, \mathbf{u}\rangle} + \min_{\mathbf{u}\neq 0} \frac{\langle \mathbf{u}, \mathbf{y}\circ\mathbf{u}\rangle}{\langle \mathbf{u}, \mathbf{u}\rangle} \\ &= \lambda_{\min}(\mathbf{x}) + \lambda_{\min}(\mathbf{y}).\end{aligned}$$

En outre,

$$\begin{aligned}\|\mathbf{y}\|_F &= \sqrt{\sum_{i=1}^{r} \|\mathbf{y}_i\|_F^2} \geq \max_{i=1,\ldots,r} \|\mathbf{y}_i\|_F \geq \max_{i=1,\ldots,r} |\min_{j=1,2} \lambda^j(\mathbf{y}_i)| \\ &\geq \max_{i=1,\ldots,r} (-\min_{j=1,2} \lambda^j(\mathbf{y}_i)) = -\min_{i=1,\ldots,r} \min_{j=1,2} \lambda^j(\mathbf{y}_i) = -\lambda_{\min}(\mathbf{y}).\end{aligned}$$

Par conséquent, $\lambda_{\min}(\mathbf{y}) \geq -\|\mathbf{y}\|_F$ et l'inégalité dans 1 est démontrée. L'inégalité dans 2 se démontre exactement de la même façon en notant que

$$\|\mathbf{y}\|_F \geq \max_{i=1,\ldots,r} \|\mathbf{y}_i\|_F \geq \max_{i=1,\ldots,r} |\max_{j=1,2} \lambda^j(\mathbf{y}_i)| \geq \max_{i=1,\ldots,r} \max_{j=1,2} \lambda^j(\mathbf{y}_i) = \lambda_{\max}(\mathbf{y}).$$

□

Lemme 5.6.4 *Soient* $\mathbf{x}, \mathbf{y} \in I\!R^n$, *tels que* $\mathbf{x} \equiv (\mathbf{x}_1, \ldots, \mathbf{x}_r)$ *et* $\mathbf{y} \equiv (\mathbf{y}_1, \ldots, \mathbf{y}_r)$ *avec* $\mathbf{x}_i, \mathbf{y}_i \in I\!R^{n_i}$. *Alors, nous avons*

$$\|\mathbf{x}\circ\mathbf{y}\|_F \leq \|\mathbf{x}\|_F\|\mathbf{y}\|_F.$$

Preuve :

1)Le résultat est vrai pour les blocs $\mathbf{x}_i$, $\mathbf{y}_i$.
Remarquons tout d'abord que puisque $Arw(\mathbf{x}_i)$ est une matrice symétrique, la norme matricielle induite par la norme euclidienne équivaut à $\lambda_{\max}(Arw(\mathbf{x}_i)) = \lambda_{\max}(\mathbf{x}_i) = \mathbf{x}_{i0} + \|\bar{\mathbf{x}}_i\|$, i.e., $\|Arw(\mathbf{x}_i)\| = \mathbf{x}_{i0} + \|\bar{\mathbf{x}}_i\|$. Cela nous permet d'obtenir $\|Arw(\mathbf{x}_i)\| \leq \sqrt{2}\|\mathbf{x}_i\|$, car

$$\begin{aligned} \|Arw(\mathbf{x}_i)\|^2 - 2\|\mathbf{x}_i\|^2 &= (\mathbf{x}_{i0} + \|\bar{\mathbf{x}}_i\|)^2 - 2(\mathbf{x}_{i0}^2 + \|\bar{\mathbf{x}}_i\|^2) \\ &= -x_{i0}^2 + 2x_{i0}\|\bar{\mathbf{x}}_i\| - \|\bar{\mathbf{x}}_i\|^2 \\ &= -(x_{i0} - \|\bar{\mathbf{x}}_i\|)^2 \leq 0. \end{aligned}$$

Par conséquent, en utilisant cette inégalité, nous arrivons au résultat attendu :

$$\begin{aligned} \|\mathbf{x}_i \circ \mathbf{y}_i\|_F &= \sqrt{2}\|\mathbf{x}_i \circ \mathbf{y}_i\| = \sqrt{2}\|Arw(\mathbf{x}_i)\mathbf{y}_i\| \\ &\leq \sqrt{2}\|Arw(\mathbf{x}_i)\|\|\mathbf{y}_i\| = \sqrt{2}\|\mathbf{x}_i\|\sqrt{2}\|\mathbf{y}_i\| \\ &= \|\mathbf{x}_i\|_F\|\mathbf{y}_i\|_F. \end{aligned}$$

2)Le résultat est vrai pour $\mathbf{x}$ et $\mathbf{y}$.

$$\begin{aligned} \|\mathbf{x} \circ \mathbf{y}\|_F^2 &= \sum_{i=1}^{r} \|\mathbf{x}_i \circ \mathbf{y}_i\|_F^2 \leq \sum_{i=1}^{r} \|\mathbf{x}_i\|_F^2\|\mathbf{y}_i\|_F^2 \\ &= \sum_{i=1}^{r} \|\mathbf{x}_i\|_F^2 \left(\sum_{j=1}^{r} \|\mathbf{y}_j\|_F^2 \delta_{ij} \right) \leq \sum_{i=1}^{r} \|\mathbf{x}_i\|_F^2 \left(\sum_{j=1}^{r} \|\mathbf{y}_j\|_F^2 \right) \\ &= \|\mathbf{x}\|_F^2\|\mathbf{y}\|_F^2. \end{aligned}$$

où δ_{ij} désigne le symbole de Kroenecker.

□

Lemme 5.6.5 *Soient* $\delta_{\mathbf{x}} = \|\widetilde{\Delta\mathbf{x}}\|_F$ *et* $\delta_{\mathbf{z}} = \|\underline{\Delta\mathbf{z}}\|_F$. *Alors, nous avons*

1. *Si* $(\mathbf{x}, \mathbf{z}) \in \mathcal{N}_F(\gamma)$, *alors* $(\mathbf{x}(\alpha), \mathbf{z}(\alpha)) \in \mathcal{N}_F(\gamma) \quad \forall \alpha \in [0, \frac{\gamma\sigma\mu}{\delta_{\mathbf{x}}\delta_{\mathbf{z}}}]$.
2. *Si* $(\mathbf{x}, \mathbf{z}) \in \mathcal{N}_2(\gamma)$, *alors* $(\mathbf{x}(\alpha), \mathbf{z}(\alpha)) \in \mathcal{N}_2(\gamma) \quad \forall \alpha \in [0, \frac{\gamma\sigma\mu}{\delta_{\mathbf{x}}\delta_{\mathbf{z}}}]$.
3. *Si* $(\mathbf{x}, \mathbf{z}) \in \mathcal{N}_{-\infty}(\gamma)$, *alors* $(\mathbf{x}(\alpha), \mathbf{z}(\alpha)) \in \mathcal{N}_{-\infty}(\gamma) \quad \forall \alpha \in [0, \frac{\gamma\sigma\mu}{\delta_{\mathbf{x}}\delta_{\mathbf{z}}}]$.

Preuve : Pour les trois assertions, nous allons montrer que le résultat est vrai pour le couple $(\widetilde{\mathbf{x}}(\alpha), \underline{\mathbf{z}}(\alpha))$, ce qui est suffisant grâce au lemme (5.4.1).
1. En partant du système (5.7) et en nous servant du fait que $\widetilde{\mathbf{x}}$ et $\underline{\mathbf{z}}$ commutent,

nous avons

$$\begin{aligned}
&\widetilde{\mathbf{x}}(\alpha) \circ \underset{\sim}{\mathbf{z}}(\alpha) - \mu(\alpha)\mathbf{e} = (\widetilde{\mathbf{x}} + \alpha\widetilde{\Delta\mathbf{x}}) \circ (\underset{\sim}{\mathbf{z}} + \alpha\underset{\sim}{\Delta\mathbf{z}}) - \mu(\alpha)\mathbf{e} \\
&= \widetilde{\mathbf{x}} \circ \underset{\sim}{\mathbf{z}} + \alpha(\widetilde{\mathbf{x}} \circ \underset{\sim}{\Delta\mathbf{z}} + \underset{\sim}{\mathbf{z}} \circ \widetilde{\Delta\mathbf{x}}) + \alpha^2\widetilde{\Delta\mathbf{x}} \circ \underset{\sim}{\Delta\mathbf{z}} - \mu(\alpha)\mathbf{e} \\
&= \widetilde{\mathbf{x}} \circ \underset{\sim}{\mathbf{z}} + \alpha\widetilde{\mathbf{x}} \circ \underset{\sim}{\mathbf{z}} - \alpha\widetilde{\mathbf{x}} \circ \underset{\sim}{\mathbf{z}} + \alpha(\widetilde{\mathbf{x}} \circ \underset{\sim}{\Delta\mathbf{z}} + \underset{\sim}{\mathbf{z}} \circ \widetilde{\Delta\mathbf{x}}) + \alpha^2\widetilde{\Delta\mathbf{x}} \circ \underset{\sim}{\Delta\mathbf{z}} \\
&-(1-\alpha)\mu\mathbf{e} - \alpha\sigma\mu\mathbf{e} \\
&= (1-\alpha)(\widetilde{\mathbf{x}} \circ \underset{\sim}{\mathbf{z}} - \mu\mathbf{e}) + \alpha\underbrace{(\widetilde{\mathbf{x}} \circ \underset{\sim}{\Delta\mathbf{z}} + \underset{\sim}{\mathbf{z}} \circ \widetilde{\Delta\mathbf{x}} + \widetilde{\mathbf{x}} \circ \underset{\sim}{\mathbf{z}} - \sigma\mu\mathbf{e})}_{=0} + \alpha^2\widetilde{\Delta\mathbf{x}} \circ \underset{\sim}{\Delta\mathbf{z}} \\
&= (1-\alpha)(\widetilde{\mathbf{w}} - \mu\mathbf{e}) + \alpha^2\widetilde{\Delta\mathbf{x}} \circ \underset{\sim}{\Delta\mathbf{z}}
\end{aligned}$$

Ensuite, en utilisant le lemme (5.6.4) et le fait que $\widetilde{\mathbf{w}}$ et $\mathbf{w}$ possèdent les mêmes valeurs propres (théorème 3.3.3), nous obtenons

$$\begin{aligned}
\|\widetilde{\mathbf{x}}(\alpha) \circ \underset{\sim}{\mathbf{z}}(\alpha) - \mu(\alpha)\mathbf{e}\|_F &\leq (1-\alpha)\|\widetilde{\mathbf{w}} - \mu\mathbf{e}\|_F + \alpha^2\|\widetilde{\Delta\mathbf{x}} \circ \underset{\sim}{\Delta\mathbf{z}}\|_F \\
&\leq (1-\alpha)\|\mathbf{w} - \mu\mathbf{e}\|_F + \alpha^2\delta_{\mathbf{x}}\delta_{\mathbf{z}} \\
&\leq (1-\alpha)\gamma\mu + \alpha^2\delta_{\mathbf{x}}\delta_{\mathbf{z}}.
\end{aligned}$$

D'où,

$$\|\widetilde{\mathbf{x}}(\alpha) \circ \underset{\sim}{\mathbf{z}}(\alpha) - \mu(\alpha)\mathbf{e}\|_F \leq \gamma\mu(\alpha)$$

si

$$(1-\alpha)\gamma\mu + \alpha^2\delta_{\mathbf{x}}\delta_{\mathbf{z}} \leq \gamma\mu(\alpha) = \gamma(1-\alpha+\sigma\alpha)\mu,$$

ou encore si

$$0 \leq \alpha \leq \frac{\gamma\sigma\mu}{\delta_{\mathbf{x}}\delta_{\mathbf{z}}} \equiv \alpha^*.$$

Par conséquent, si nous parvenons à montrer que dans $[0,\ \alpha^*]$ $\widetilde{\mathbf{x}}(\alpha)$ et $\underset{\sim}{\mathbf{z}}(\alpha)$ sont définis positifs le point 1 sera démontré grâce à l'inégalité (5.9) du lemme (5.6.2).

Puisque $\mu(\alpha) - \lambda_{\min}(\widetilde{\mathbf{x}}(\alpha) \circ \underset{\sim}{\mathbf{z}}(\alpha)) \leq \|\widetilde{\mathbf{x}}(\alpha) \circ \underset{\sim}{\mathbf{z}}(\alpha) - \mu(\alpha)\mathbf{e}\|_F$, nous aurons dans $[0,\ \alpha^*]$, $\mu(\alpha) - \lambda_{\min}(\widetilde{\mathbf{x}}(\alpha) \circ \underset{\sim}{\mathbf{z}}(\alpha)) \leq \gamma\mu(\alpha)$ et donc

$$\lambda_{\min}(\widetilde{\mathbf{x}}(\alpha) \circ \underset{\sim}{\mathbf{z}}(\alpha)) \geq \mu(\alpha)(1-\gamma) > 0. \tag{5.12}$$

Par définition de $\lambda_{\min}$, nous avons que pour $\alpha \in [0, \alpha^*]$ et pour $i = 1, \ldots, r$, $\widetilde{\mathbf{x}}_i(\alpha) \circ \underset{\sim}{\mathbf{z}}_i(\alpha)$ est défini positif. Par conséquent, dans $[0,\ \alpha^*]$, nous avons

$$\det(\widetilde{\mathbf{x}}(\alpha) \circ \underset{\sim}{\mathbf{z}}(\alpha)) = \prod_{i=1}^{r} \det(\widetilde{\mathbf{x}}_i(\alpha) \circ \underset{\sim}{\mathbf{z}}_i(\alpha)) > 0.$$

Nous allons supposer par l'absurde que $\widetilde{\mathbf{x}}(\alpha)$, par exemple, n'est pas défini positif dans $[0, \alpha^*]$, c'est-à-dire que $\lambda_{\min}(\widetilde{\mathbf{x}}(\alpha))$ prend une valeur négative ou nulle dans cet intervalle pour un certain α. Puisque par hypothèse $\mathbf{x}$ (et donc $\widetilde{\mathbf{x}}$) est défini positif, il suit que $\lambda_{\min}(\widetilde{\mathbf{x}}) = \lambda_{\min}(\widetilde{\mathbf{x}}(0)) > 0$. Mais par continuité de la valeur propre $\lambda_{\min}(\widetilde{\mathbf{x}}(\cdot))$, il doit nécessairement exister un certain $\hat{\alpha} \in [0, \alpha^*]$ tq $\lambda_{\min}(\widetilde{\mathbf{x}}(\hat{\alpha})) = 0$ (si il existe plusieurs valeurs de α dans $[0, \alpha^*]$ telles que

$\lambda_{\min}(\widetilde{\mathbf{x}}(\alpha)) = 0$, alors $\hat{\alpha}$ désignera la plus petite de celles-ci). Avec cet $\hat{\alpha}$, nous obtenons par la propriété 4 du théorème (3.3.1) que

$$\det(\widetilde{\mathbf{w}}(\hat{\alpha})) = \det(Q_{(\widetilde{\mathbf{x}}(\hat{\alpha}))^{1/2}\underset{\sim}{\mathbf{z}}(\hat{\alpha})}) = \det(\widetilde{\mathbf{x}}(\hat{\alpha}))\det(\underset{\sim}{\mathbf{z}}(\hat{\alpha})) = 0. \tag{5.13}$$

Ensuite, à nouveau par continuité de $\lambda_{\min}(\widetilde{\mathbf{x}}(\cdot))$, nous avons que pour un $\varepsilon > 0$ suffisament petit $\lambda_{\min}(\widetilde{\mathbf{x}}(\hat{\alpha}-\varepsilon)) > 0$, ce qui revient à dire que $\widetilde{\mathbf{x}}(\hat{\alpha}-\varepsilon)$ est défini positif. Donc, grâce à la troisième inégalité du lemme (5.6.2) qui est obtenue sans supposer que z soit défini positif nous obtenons

$$\lambda_{\min}(\widetilde{\mathbf{w}}(\hat{\alpha}-\varepsilon)) \geq \lambda_{\min}(\widetilde{\mathbf{x}}(\hat{\alpha}-\varepsilon) \circ \underset{\sim}{\mathbf{z}}(\hat{\alpha}-\varepsilon)).$$

En passant à la limite lorsque $\varepsilon \to 0$ et en utilisant l'inégalité dans (5.12), nous arrivons à

$$\lambda_{\min}(\widetilde{\mathbf{w}}(\hat{\alpha})) \geq \lambda_{\min}(\widetilde{\mathbf{x}}(\hat{\alpha}) \circ \underset{\sim}{\mathbf{z}}(\hat{\alpha})) > 0$$

qui entre en contradiction avec (5.13). D'où, $\widetilde{\mathbf{x}}(\alpha)$ doit être défini positif dans $[0, \alpha^*]$. En supposant que $\underset{\sim}{\mathbf{z}}(\alpha)$ soit non défini positif, nous arrivons de façon tout-à-fait similaire à la même contradiction.

3. Pour montrer cette assertion, il nous suffit de montrer que $\widetilde{\mathbf{x}}(\alpha)$ et $\underset{\sim}{\mathbf{z}}(\alpha)$ sont définis positifs et que

$$\mu(\alpha) - \lambda_{\min}(\widetilde{\mathbf{x}}(\alpha) \circ \underset{\sim}{\mathbf{z}}(\alpha)) \leq \gamma\mu(\alpha) \quad \forall\, \alpha \in [0, \frac{\gamma\sigma\mu}{\delta_{\mathbf{x}}\delta_{\mathbf{z}}}],$$

car alors, en vertu de l'inégalité (5.11) du lemme (5.6.2), nous obtiendrons $(\widetilde{\mathbf{x}}(\alpha), \underset{\sim}{\mathbf{z}}(\alpha)) \in \mathcal{N}_{-\infty}(\gamma)$.
Puisque $(\mathbf{x}, \mathbf{z}) \in \mathcal{N}_{-\infty}$ nous avons $\mu - \lambda_{\min}(\mathbf{w}) \leq \gamma\mu$ ou encore $\lambda_{\min}(\mathbf{w}) - \mu \geq -\gamma\mu$. En utilisant le fait que $\widetilde{\mathbf{x}}$ et $\underset{\sim}{\mathbf{z}}$ commutent ainsi que le lemme (5.6.3), il suit que

$$\begin{aligned}
\lambda_{\min}(\widetilde{\mathbf{x}}(\alpha) \circ \underset{\sim}{\mathbf{z}}(\alpha) - \mu(\alpha)\mathbf{e}) &= \lambda_{\min}((1-\alpha)(\widetilde{\mathbf{w}} - \mu\mathbf{e}) + \alpha^2 \widetilde{\Delta\mathbf{x}} \circ \underset{\sim}{\Delta\mathbf{z}}) \\
&\geq (1-\alpha)\lambda_{\min}(\widetilde{\mathbf{w}} - \mu\mathbf{e}) - \alpha^2 \|\widetilde{\Delta\mathbf{x}} \circ \underset{\sim}{\Delta\mathbf{z}}\|_F \\
&\geq (1-\alpha)\lambda_{\min}(\widetilde{\mathbf{w}} - \mu\mathbf{e}) - \alpha^2\delta_{\mathbf{x}}\delta_{\mathbf{z}} \\
&= (1-\alpha)\lambda_{\min}(\mathbf{w} - \mu\mathbf{e}) - \alpha^2\delta_{\mathbf{x}}\delta_{\mathbf{z}} \\
&\geq (\alpha-1)\gamma\mu - \alpha^2\delta_{\mathbf{x}}\delta_{\mathbf{z}}.
\end{aligned}$$

Donc,

$$\mu(\alpha) - \lambda_{\min}(\widetilde{\mathbf{x}}(\alpha) \circ \underset{\sim}{\mathbf{z}}(\alpha)) \leq \gamma\mu(\alpha)$$

si

$$(1-\alpha)\gamma\mu + \alpha^2\delta_{\mathbf{x}}\delta_{\mathbf{z}} \leq \gamma\mu(\alpha) = \gamma\mu(1-\alpha) + \gamma\mu\sigma\alpha$$

c'est-à-dire si $0 \leq \alpha \leq \frac{\gamma\sigma\mu}{\delta_{\mathbf{x}}\delta_{\mathbf{z}}}$. Dès lors, dans l'intervalle $[0, \frac{\gamma\sigma\mu}{\delta_{\mathbf{x}}\delta_{\mathbf{z}}}]$, l'inégalité (5.12) est vérifiée. De la même manière qu'en 1, nous établissons que $\widetilde{\mathbf{x}}(\alpha)$ et $\underset{\sim}{\mathbf{z}}(\alpha)$ sont définis positifs et le résultat annoncé est démontré.

2. Ici aussi, il nous suffit de montrer que $\|\widetilde{\mathbf{x}}(\alpha) \circ \underset{\sim}{\mathbf{z}}(\alpha) - \mu(\alpha)\mathbf{e}\|_2 \leq \gamma\mu(\alpha)$ et que $\widetilde{\mathbf{x}}(\alpha)$ et $\underset{\sim}{\mathbf{z}}(\alpha)$ sont définis positifs dans $[0, \alpha^*]$. Notons tout d'abord que puisque $\mathcal{N}_2(\gamma) \subseteq \mathcal{N}_{-\infty}(\gamma)$, nous avons comme dans la preuve de 3 que dans l'intervalle $[0, \alpha^*]$, $\mu(\alpha) - \lambda_{\min}(\widetilde{\mathbf{x}}(\alpha) \circ \underset{\sim}{\mathbf{z}}(\alpha)) \leq \gamma\mu(\alpha)$ ou encore

$$\lambda_{\max}(\mu(\alpha)\mathbf{e} - \widetilde{\mathbf{x}}(\alpha) \circ \underset{\sim}{\mathbf{z}}(\alpha)) \leq \gamma\mu(\alpha). \tag{5.14}$$

En utilisant les lemmes (5.6.3) et (5.6.4) ainsi que le fait que $\widetilde{\mathbf{x}}$ et $\underset{\sim}{\mathbf{z}}$ commutent, nous obtenons

$$\begin{aligned}
\lambda_{\max}(\widetilde{\mathbf{x}}(\alpha) \circ \underset{\sim}{\mathbf{z}}(\alpha) - \mu(\alpha)\mathbf{e}) &= \lambda_{\max}((1-\alpha)(\widetilde{\mathbf{w}} - \mu\mathbf{e}) + \alpha^2 \widetilde{\Delta\mathbf{x}} \circ \underset{\sim}{\Delta\mathbf{z}}) \\
&\leq (1-\alpha)\lambda_{\max}(\widetilde{\mathbf{w}} - \mu\mathbf{e}) + \alpha^2 \|\widetilde{\Delta\mathbf{x}} \circ \underset{\sim}{\Delta\mathbf{z}}\|_F \\
&\leq (1-\alpha)\lambda_{\max}(\mathbf{w} - \mu\mathbf{e}) + \alpha^2 \delta_{\mathbf{x}}\delta_{\mathbf{z}} \\
&\leq (1-\alpha)\mu\gamma + \alpha^2 \delta_{\mathbf{x}}\delta_{\mathbf{z}}
\end{aligned}$$

donc, si

$$(1-\alpha)\gamma\mu + \alpha^2 \delta_{\mathbf{x}}\delta_{\mathbf{z}} \leq \gamma\mu(\alpha)$$

i.e., si $0 \leq \alpha \leq \alpha^* \equiv \frac{\gamma\sigma\mu}{\delta_{\mathbf{x}}\delta_{\mathbf{z}}}$, nous aurons

$$\lambda_{\max}(\widetilde{\mathbf{x}}(\alpha) \circ \underset{\sim}{\mathbf{z}}(\alpha) - \mu(\alpha)\mathbf{e}) \leq \gamma\mu(\alpha). \tag{5.15}$$

Le caractère défini positif de $\widetilde{\mathbf{x}}(\alpha)$ et $\underset{\sim}{\mathbf{z}}(\alpha)$ est obtenu comme dans le point 3 en se servant de l'inégalité $\mu(\alpha) - \lambda_{\min}(\widetilde{\mathbf{x}}(\alpha) \circ \underset{\sim}{\mathbf{z}}(\alpha)) \leq \gamma\mu(\alpha)$ et en raisonnant de la même manière qu'en 1. Finalement, en prenant le maximum dans les inégalités (5.14) et (5.15) lorsque $\alpha \in [0, \alpha^*]$, nous obtenons

$$\|\widetilde{\mathbf{x}}(\alpha) \circ \underset{\sim}{\mathbf{z}}(\alpha) - \mu(\alpha)\mathbf{e}\|_2 \leq \gamma\mu(\alpha)$$

ce qui termine la preuve de 2.

□

5.6.2 Bornes sur $\delta_{\mathbf{x}}$ et $\delta_{\mathbf{z}}$

Lemme 5.6.6 *Soient* $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ *et* G *une matrice symétrique définie positive de dimension* n. *Alors, nous avons*

$$\|\mathbf{x}\|\|\mathbf{y}\| \leq \frac{1}{2}\sqrt{cond(G)}(\|G^{1/2}\mathbf{x}\|^2 + \|G^{-1/2}\mathbf{y}\|^2), \tag{5.16}$$

où $cond(G) = \lambda_{\max}(G)/\lambda_{\min}(G)$.

Preuve : Puisque G est une matrice symétrique, nous pouvons écrire

$$\lambda_{\min}(G) = \min_{\mathbf{x}\neq 0} \frac{\langle \mathbf{x}, G\mathbf{x}\rangle}{\langle \mathbf{x}, \mathbf{x}\rangle} = \min_{\mathbf{x}\neq 0} \frac{\langle G^{1/2}\mathbf{x}, G^{1/2}\mathbf{x}\rangle}{\langle \mathbf{x}, \mathbf{x}\rangle} = \min_{\mathbf{x}\neq 0} \frac{\|G^{1/2}\mathbf{x}\|^2}{\|\mathbf{x}\|^2},$$

ce qui implique que $\forall \mathbf{x} \neq \mathbf{0} \in I\!R^n, \quad \lambda_{\min}(G) \leq \frac{\|G^{1/2}\mathbf{x}\|^2}{\|\mathbf{x}\|^2}$ ou encore

$$\forall \mathbf{x} \in I\!R^n, \quad \|\mathbf{x}\| \leq \frac{\|G^{1/2}\mathbf{x}\|}{\sqrt{\lambda_{\min}(G)}}.$$

Ainsi, en choisissant les matrices G pour $\mathbf{x}$ et G^{-1} pour $\mathbf{y}$, nous obtenons

$$\|\mathbf{x}\|\|\mathbf{y}\| \leq \frac{\|G^{1/2}\mathbf{x}\|}{\sqrt{\lambda_{\min}(G)}} \frac{\|G^{-1/2}\mathbf{y}\|}{\sqrt{\lambda_{\min}(G^{-1})}} = \sqrt{\frac{1}{\lambda_{\min}(G)\lambda_{\min}(G^{-1})}} \|G^{1/2}\mathbf{x}\|\|G^{-1/2}\mathbf{y}\|.$$

Par conséquent, en notant que $\lambda_{\min}(G^{-1}) = 1/\lambda_{\max}(G)$ et que pour des réels a, b : $2ab \leq a^2 + b^2$, nous arrivons au résultat annoncé. Remarquons que le résultat reste vrai pour la norme de Frobénius car $\|\cdot\|_F = \sqrt{2}\|\cdot\|$.

□

Lemme 5.6.7 *Soient* $\mathbf{x}$, $\mathbf{z} \succ_Q \mathbf{0}$ *et soit* $G = Arw^{-1}(\underset{\sim}{\mathbf{z}})Arw(\widetilde{\mathbf{x}})$. *Alors,*

$$\delta_{\mathbf{x}}\delta_{\mathbf{z}} \leq \frac{1}{2}\sqrt{cond(G)} \sum_{\substack{i=1,\ldots,r \\ j=1,2}} \frac{(\sigma\mu - \lambda^j(\widetilde{\mathbf{w}}_i))^2}{\lambda^j(\widetilde{\mathbf{w}}_i)}.$$

Preuve : Puisque $\widetilde{\mathbf{x}}$ et $\underset{\sim}{\mathbf{z}}$ commutent, il en est de même pour les matrices symétriques définies positives $Arw(\widetilde{\mathbf{x}})$ et $Arw(\underset{\sim}{\mathbf{z}})$, ce qui rend la matrice G symétrique. De plus, G est définie positive puisque

$$Arw^{-1}(\underset{\sim}{\mathbf{z}})Arw(\widetilde{\mathbf{x}}) = Arw^{-1/2}(\underset{\sim}{\mathbf{z}})Arw^{1/2}(\widetilde{\mathbf{x}})Arw^{1/2}(\widetilde{\mathbf{x}})Arw^{-1/2}(\underset{\sim}{\mathbf{z}}).$$

Par conséquent, en utilisant le lemme (5.6.6), nous obtenons

$$\delta_{\mathbf{x}}\delta_{\mathbf{z}} \leq \frac{1}{2}\sqrt{cond(G)}(\|G^{1/2}\underset{\sim}{\Delta\mathbf{z}}\|_F^2 + \|G^{-1/2}\widetilde{\Delta\mathbf{x}}\|_F^2).$$

Donc, pour obtenir le résultat il nous suffit de montrer que

$$\|G^{1/2}\underset{\sim}{\Delta\mathbf{z}}\|_F^2 + \|G^{-1/2}\widetilde{\Delta\mathbf{x}}\|_F^2 = \sum_{\substack{i=1,\ldots,r \\ j=1,2}} \frac{(\sigma\mu - \lambda^j(\widetilde{\mathbf{w}}_i))^2}{\lambda^j(\widetilde{\mathbf{w}}_i)}.$$

Nous avons

$$(Arw(\widetilde{\mathbf{x}})Arw(\underset{\sim}{\mathbf{z}}))^{-1/2}Arw(\widetilde{\mathbf{x}}) = Arw^{-1/2}(\underset{\sim}{\mathbf{z}})Arw^{1/2}(\widetilde{\mathbf{x}}) = G^{1/2}$$
$$(Arw(\widetilde{\mathbf{x}})Arw(\underset{\sim}{\mathbf{z}}))^{-1/2}Arw(\underset{\sim}{\mathbf{z}}) = Arw^{1/2}(\underset{\sim}{\mathbf{z}})Arw^{-1/2}(\widetilde{\mathbf{x}}) = G^{-1/2}.$$

Cela implique qu'en pré-multipliant la dernière équation de (5.7) par $(Arw(\widetilde{\mathbf{x}})Arw(\underset{\sim}{\mathbf{z}}))^{-1/2}$, nous obtenons l'égalité suivante

$$G^{1/2}\underset{\sim}{\Delta \mathbf{z}} + G^{-1/2}\widetilde{\Delta \mathbf{x}} = \sigma\mu(Arw(\widetilde{\mathbf{x}})Arw(\underset{\sim}{\mathbf{z}}))^{-1/2}\mathbf{e} - G^{1/2}\underset{\sim}{\mathbf{z}}.$$

De plus,

$$\begin{aligned}\|G^{1/2}\underset{\sim}{\Delta \mathbf{z}}\|_F^2 + \|G^{-1/2}\widetilde{\Delta \mathbf{x}}\|_F^2 &= 2(\|G^{1/2}\underset{\sim}{\Delta \mathbf{z}}\|^2 + \|G^{-1/2}\widetilde{\Delta \mathbf{x}}\|^2) \\ &= 2\|G^{1/2}\underset{\sim}{\Delta \mathbf{z}} + G^{-1/2}\widetilde{\Delta \mathbf{x}}\|^2\end{aligned}$$

où la seconde égalité est obtenue grâce au point 1 du lemme (5.6.1). Nous obtenons alors,

$$\begin{aligned}&\|G^{1/2}\underset{\sim}{\Delta \mathbf{z}} + G^{-1/2}\widetilde{\Delta \mathbf{x}}\|^2 = \|\sigma\mu(Arw(\widetilde{\mathbf{x}})Arw(\underset{\sim}{\mathbf{z}}))^{-1/2}\mathbf{e} - G^{1/2}\underset{\sim}{\mathbf{z}}\|^2 \\ &= \sigma^2\mu^2\langle(Arw(\widetilde{\mathbf{x}})Arw(\underset{\sim}{\mathbf{z}}))^{-1/2}\mathbf{e},\ (Arw(\widetilde{\mathbf{x}})Arw(\underset{\sim}{\mathbf{z}}))^{-1/2}\mathbf{e}\rangle \\ &\quad -2\sigma\mu\langle(Arw(\widetilde{\mathbf{x}})Arw(\underset{\sim}{\mathbf{z}}))^{-1/2}\mathbf{e},\ (Arw(\widetilde{\mathbf{x}})Arw(\underset{\sim}{\mathbf{z}}))^{-1/2}Arw(\widetilde{\mathbf{x}})\underset{\sim}{\mathbf{z}}\rangle \\ &\quad +\langle G^{1/2}\underset{\sim}{\mathbf{z}},\ G^{1/2}\underset{\sim}{\mathbf{z}}\rangle \\ &= \sigma^2\mu^2\langle\mathbf{e},\ (Arw(\widetilde{\mathbf{x}})Arw(\underset{\sim}{\mathbf{z}}))^{-1}\mathbf{e}\rangle - 2\sigma\mu\langle(Arw(\widetilde{\mathbf{x}})Arw(\underset{\sim}{\mathbf{z}}))^{-1}\mathbf{e},\ Arw(\widetilde{\mathbf{x}})\underset{\sim}{\mathbf{z}}\rangle \\ &\quad +\langle G\underset{\sim}{\mathbf{z}},\ \underset{\sim}{\mathbf{z}}\rangle \\ &= \sigma^2\mu^2\langle Arw(\widetilde{\mathbf{x}})\widetilde{\mathbf{x}}^{-1},\ (Arw(\widetilde{\mathbf{x}})Arw(\underset{\sim}{\mathbf{z}}))^{-1}Arw(\underset{\sim}{\mathbf{z}})\underset{\sim}{\mathbf{z}}^{-1}\rangle \\ &\quad -2\sigma\mu\langle(Arw(\widetilde{\mathbf{x}})Arw(\underset{\sim}{\mathbf{z}}))^{-1}\mathbf{e},\ Arw(\widetilde{\mathbf{x}})Arw(\underset{\sim}{\mathbf{z}})\mathbf{e}\rangle + \langle Arw^{-1}(\underset{\sim}{\mathbf{z}})Arw(\widetilde{\mathbf{x}})\underset{\sim}{\mathbf{z}}, \\ &\quad Arw(\underset{\sim}{\mathbf{z}})\mathbf{e}\rangle \\ &= \sigma^2\mu^2\langle\widetilde{\mathbf{x}}^{-1},\ \underset{\sim}{\mathbf{z}}^{-1}\rangle - 2\sigma\mu\langle\mathbf{e},\ \mathbf{e}\rangle + \langle\widetilde{\mathbf{x}}\circ\underset{\sim}{\mathbf{z}},\ \mathbf{e}\rangle \\ &= \sigma^2\mu^2\langle\mathbf{e},\ \widetilde{\mathbf{x}}^{-1}\circ\underset{\sim}{\mathbf{z}}^{-1}\rangle - 2\sigma\mu\langle\mathbf{e},\ \mathbf{e}\rangle + \langle\widetilde{\mathbf{x}}\circ\underset{\sim}{\mathbf{z}},\ \mathbf{e}\rangle \\ &= \sigma^2\mu^2\langle\mathbf{e},\ \widetilde{\mathbf{w}}^{-1}\rangle - 2\sigma\mu\langle\mathbf{e},\ \mathbf{e}\rangle + \langle\widetilde{\mathbf{w}},\ \mathbf{e}\rangle \text{ (car } \widetilde{\mathbf{x}} \text{ et } \underset{\sim}{\mathbf{z}} \text{ commutent)} \\ &= \tfrac{1}{2}(\sigma^2\mu^2 tr(\widetilde{\mathbf{w}}^{-1}) - 2\sigma\mu tr(\mathbf{e}) + tr(\widetilde{\mathbf{w}})) \\ &= \tfrac{1}{2}(\sigma^2\mu^2\sum_{\substack{i=1,\dots,r\\ j=1,2}}\frac{1}{\lambda^j(\widetilde{\mathbf{w}}_i)} - 2\sigma\mu\sum_{\substack{i=1,\dots,r\\ j=1,2}}1 + \sum_{\substack{i=1,\dots,r\\ j=1,2}}\lambda^j(\widetilde{\mathbf{w}}_i)) \\ &= \frac{1}{2}\sum_{\substack{i=1,\dots,r\\ j=1,2}}\frac{(\sigma\mu - \lambda^j(\widetilde{\mathbf{w}}_i))^2}{\lambda^j(\widetilde{\mathbf{w}}_i)},\end{aligned}$$

ce qui termine la preuve.

□

5.6.3 Bornes sur le nombre de conditionnement de G

Lemme 5.6.8 *Soient* $\mathbf{x}, \mathbf{z} \succ_{\mathcal{Q}} \mathbf{0}$ *et soit* $G = Arw^{-1}(\underset{\sim}{z})Arw(\tilde{\mathbf{x}})$. *Alors, pour chacun des voisinages, nous avons les bornes suivantes pour* $\sum_{\substack{i=1,\ldots,r \\ j=1,2}}(\sigma\mu - \lambda^j(\tilde{\mathbf{w}}_i))^2/\lambda^j(\tilde{\mathbf{w}}_i)$:

1. *Si* $(\mathbf{x}, \mathbf{z}) \in \mathcal{N}_F(\gamma)$, *alors*

$$\sum_{\substack{i=1,\ldots,r \\ j=1,2}} \frac{(\sigma\mu - \lambda^j(\tilde{\mathbf{w}}_i))^2}{\lambda^j(\tilde{\mathbf{w}}_i)} \leq \left(\frac{\gamma^2 + (1-\sigma)^2 2r}{1-\gamma}\right)\mu.$$

2. *Si* $(\mathbf{x}, \mathbf{y}) \in \mathcal{N}_2(\gamma)$ *ou* $\mathcal{N}_{-\infty}(\gamma)$, *alors*

$$\sum_{\substack{i=1,\ldots,r \\ j=1,2}} \frac{(\sigma\mu - \lambda^j(\tilde{\mathbf{w}}_i))^2}{\lambda^j(\tilde{\mathbf{w}}_i)} \leq \left(1 - 2\sigma + \frac{\sigma^2}{1-\gamma}\right) 2r\mu.$$

Preuve : 1) Puisque $\mathcal{N}_F(\gamma) \subseteq \mathcal{N}_{-\infty}(\gamma)$ et $(\tilde{\mathbf{x}}, \underset{\sim}{z}) \in \mathcal{N}_F(\gamma)$, nous avons $\mu - \lambda_{\min}(\tilde{\mathbf{w}}) \leq \gamma\mu$, i.e., $\lambda_{\min}(\tilde{\mathbf{w}}) \geq (1-\gamma)\mu$. Par conséquent, il suit que

$$\begin{aligned}
\sum_{\substack{i=1,\ldots,r \\ j=1,2}} \frac{(\sigma\mu - \lambda^j(\tilde{\mathbf{w}}_i))^2}{\lambda^j(\tilde{\mathbf{w}}_i)} &\leq \tfrac{1}{(1-\gamma)\mu} \sum_{\substack{i=1,\ldots,r \\ j=1,2}} (\sigma\mu - \lambda^j(\tilde{\mathbf{w}}_i))^2 \\
&= \tfrac{1}{(1-\gamma)\mu} \sum_{\substack{i=1,\ldots,r \\ j=1,2}} (\mu - \lambda^j(\tilde{\mathbf{w}}_i) - (1-\sigma)\mu)^2 \\
&\leq \tfrac{1}{(1-\gamma)\mu} \sum_{\substack{i=1,\ldots,r \\ j=1,2}} (\mu - \lambda^j(\tilde{\mathbf{w}}_i))^2 + ((1-\sigma)\mu)^2 \\
&\leq \tfrac{1}{(1-\gamma)\mu}(\gamma^2\mu^2 + 2r(1-\sigma)^2\mu^2), \text{ (car } (\tilde{\mathbf{x}}, \underset{\sim}{z}) \in \mathcal{N}_F(\gamma)) \\
&\leq \left(\tfrac{\gamma^2 + 2r(1-\sigma)^2}{1-\gamma}\right)\mu,
\end{aligned}$$

où la seconde inégalité est obtenue en remarquant que

$$\begin{aligned}
-2 \sum_{\substack{i=1,\ldots,r \\ j=1,2}} (\mu - \lambda^j(\tilde{\mathbf{w}}_i))(1-\sigma)\mu &= -2(1-\sigma)\mu(2r\mu - \sum_{\substack{i=1,\ldots,r \\ j=1,2}} \lambda^j(\tilde{\mathbf{w}}_i)) \\
&\leq -2(1-\sigma)\mu(2r\mu + \sum_{\substack{i=1,\ldots,r \\ j=1,2}} (\gamma - 1)\mu) \\
&= -2(1-\sigma)\mu(2r\mu + 2r\gamma\mu - 2r\mu) \\
&= -4r(1-\sigma)\mu^2\gamma \leq 0.
\end{aligned}$$

2) Que $(\mathbf{x}, \mathbf{z})$ soit dans $\mathcal{N}_2(\gamma)$ ou dans $\mathcal{N}_{-\infty}(\gamma)$, nous avons que $(\tilde{\mathbf{x}}, \underset{\sim}{z}) \in \mathcal{N}_{-\infty}(\gamma)$ ce qui nous permet de minorer toutes les valeurs propres de $\tilde{\mathbf{w}}$ par $(1-\gamma)\mu$.

Puisque les valeurs propres de $\widetilde{\mathbf{w}}^{-1}$ sont les inverses des valeurs propes de $\widetilde{\mathbf{w}}$, nous pouvons les majorer par $1/(1-\gamma)\mu$. Par conséquent, nous obtenons :

$$\begin{aligned}
\sum_{\substack{i=1,\ldots,r \\ j=1,2}} \frac{(\sigma\mu - \lambda^j(\widetilde{\mathbf{w}}_i))^2}{\lambda^j(\widetilde{\mathbf{w}}_i)} &= \sigma^2\mu^2 \sum_{\substack{i=1,\ldots,r \\ j=1,2}} (\lambda^j(\widetilde{\mathbf{w}}_i^{-1})) - 4\sigma\mu r + tr(\widetilde{\mathbf{w}}) \\
&\leq \sigma^2\mu^2 \tfrac{2r}{(1-\gamma)\mu} - 4\sigma\mu r + tr(\widetilde{\mathbf{x}} \circ \underset{\sim}{\mathbf{z}}) \\
&= \sigma^2\mu \tfrac{2r}{(1-\gamma)} - 4\sigma\mu r + 2\langle \widetilde{\mathbf{x}}, \underset{\sim}{\mathbf{z}} \rangle \\
&= \sigma^2\mu \tfrac{2r}{(1-\gamma)} - 4\sigma\mu r + 2\mu r \\
&= \left(1 - 2\sigma + \tfrac{\sigma^2}{1-\gamma}\right) 2\mu r
\end{aligned}$$

ce qui termine la preuve de 2.

□

Lemme 5.6.9 *Pour la méthode de Nesterov-Todd, le conditionnement de G vaut toujours 1. Pour les méthodes* **xz** *et* **zx**, *nous avons*

1. *Si* $(\mathbf{x}, \mathbf{z})$ *est dans* $\mathcal{N}_F(\gamma)$ *ou* $\mathcal{N}_2(\gamma)$ *alors* $cond(G) \leq 2/(1-\gamma)$.
2. *Si* $(\mathbf{x}, \mathbf{z})$ *est dans* $\mathcal{N}_{-\infty}(\gamma)$ *alors* $cond(G) \leq 2r/(1-\gamma)$.

Preuve : Dans la méthode de Nesterov-Todd, $\mathbf{p}$ est choisi de sorte que $\widetilde{\mathbf{x}} = \underset{\sim}{\mathbf{z}}$. D'où, $G = Arw^{-1}(\underset{\sim}{\mathbf{z}})Arw(\widetilde{\mathbf{x}}) = I$, et de manière évidente $cond(G) = 1$.
Dans la méthode **xz**, $\mathbf{p}$ est choisi tel que $\underset{\sim}{\mathbf{z}} = \mathbf{e}$. Donc, $\widetilde{\mathbf{x}} = \widetilde{\mathbf{w}}$ et $G = Arw(\widetilde{\mathbf{w}})$.

1) Si $(\mathbf{x}, \mathbf{z}) \in \mathcal{N}_2(\gamma)$, il en est de même pour $(\widetilde{\mathbf{x}}, \underset{\sim}{\mathbf{z}})$, ce qui nous permet de majorer $\lambda_{\max}(\widetilde{\mathbf{w}})(= \lambda_{\max}(G))$ par $\mu(1+\gamma)$. De plus, puisque $(\widetilde{\mathbf{x}}, \underset{\sim}{\mathbf{z}})$ appartient aussi à $\mathcal{N}_{-\infty}(\gamma)$, $\lambda_{\min}(\widetilde{\mathbf{w}})(= \lambda_{\min}(G))$ peut être minorée par $(1-\gamma)\mu$. D'où,

$$cond(G) = \frac{\lambda_{\max}(G)}{\lambda_{\min}(G)} \leq \frac{1+\gamma}{1-\gamma} \leq \frac{2}{1-\gamma}.$$

Etant donné que $\mathcal{N}_F(\gamma) \subseteq \mathcal{N}_2(\gamma)$, ceci termine la preuve de 1 pour la méthode **xz**.

2) Lorsque $(\mathbf{x}, \mathbf{z}) \in \mathcal{N}_{-\infty}(\gamma)$, nous pouvons utiliser la même borne inférieure pour $\lambda_{\min}(\widetilde{\mathbf{w}})$ que celle utilisée dans la preuve de 1, et $\lambda_{\max}(\widetilde{\mathbf{w}})$ peut être de manière évidente majorée par $tr(\widetilde{\mathbf{w}}) = 2\langle \widetilde{\mathbf{x}}, \underset{\sim}{\mathbf{z}} \rangle = 2\mu r$. Donc, nous aboutissons à

$$cond(G) \leq \frac{2\mu r}{(1-\gamma)\mu} = \frac{2r}{1-\gamma},$$

ce qui prouve le point 2 pour la méthode **xz**.
Dans la méthode **zx**, nous avons $\widetilde{\mathbf{x}} = \mathbf{e}$ et donc $\underset{\sim}{\mathbf{z}} = \widetilde{\mathbf{w}}$. D'où, $G = Arw^{-1}(\widetilde{\mathbf{w}})$ et puisqu'une matrice et son inverse ont le même nombre de conditionnement, la preuve est achevée.

□

5.6.4 Complexité des algorithmes

Les algorithmes de suivi de chemins à petits pas, à pas semi-longs et à longs pas pour la classe commutative des directions peuvent être décrits de la façon suivante :

Etape 1 :
- Choisir ϵ, σ, $\gamma \in (0,1)$ et $(\mathbf{x}^0, y^0, \mathbf{z}^0) \in \mathcal{N}_\bullet(\gamma)$ pour un des trois voisinages définis.
- Poser $k = 0$ et $\mu^0 = \langle \mathbf{x}^0, \mathbf{z}^0 \rangle / r$.

Etape 2 : Tant que $\mu^k > \epsilon\mu_0$ répeter
- Choisir un $\mathbf{p} \in \mathcal{C}(\mathbf{x}^k, \mathbf{z}^k)$ et calculer $(\widetilde{\mathbf{x}}^k, y^k, \underset{\sim}{\mathbf{z}}^k)$.
- Calculer la direction de Newton $(\widetilde{\Delta\mathbf{x}}^k, \Delta y^k, \underset{\sim}{\Delta\mathbf{z}}^k)$ en résolvant le système (5.7) et appliquer le changement d'échelle inverse pour obtenir $(\Delta\mathbf{x}^k, \Delta y^k, \Delta\mathbf{z}^k)$.
- Déterminer α_k comme étant la longueur de pas maximale tq

$$(\mathbf{x}^{k+1}, y^{k+1}, \mathbf{z}^{k+1}) = (\mathbf{x}^k, y^k, \mathbf{z}^k) + \alpha_k(\Delta\mathbf{x}^k, \Delta y^k, \Delta\mathbf{z}^k) \in \mathcal{N}_\bullet(\gamma).$$

- Poser $\mu^{k+1} = \langle \mathbf{x}^{k+1}, \mathbf{z}^{k+1} \rangle / r$ et incrémenter k.

Dans cet algorithme, le paramètre ϵ est la tolérance choisie pour la solution finale. Le choix de σ, γ et du voisinage détermine le type général de l'algorithme. En choisissant $\mathcal{N}_F(\gamma)$ et $\sigma = 1 - \delta/\sqrt{r}$, où $\delta \in (0, 1)$ nous obtenons l'algorithme à petit pas. L'algorithme à pas semi-longs est obtenu avec $\mathcal{N}_2(\gamma)$ et $\sigma \in (0, 1)$. Finalement, en choisissant $\mathcal{N}_{-\infty}(\gamma)$ et $\sigma \in (0, 1)$ nous obtenons l'algorithme à longs pas.
Notre résultat principal est résumé dans le théorème suivant qui décrit la complexité des trois algorithmes de suivi de chemins que nous venons de décrire.

Théorème 5.6.1 *Supposons que $\sqrt{cond(G)}$ peut être borné supérieurement par $\kappa < \infty$ à chaque itération de l'algorithme. Alors, l'algorithme à petit pas se termine en $\mathcal{O}(\kappa\sqrt{r}\log\epsilon^{-1})$ itérations. Les algorithmes à pas semi-longs et à longs pas se terminent en $\mathcal{O}(\kappa r \log\epsilon^{-1})$ itérations.*

Preuve : Si nous choisissons le voisinage $\mathcal{N}_F(\gamma)$ et $\sigma = 1 - \delta/\sqrt{r}$, alors en vertu des lemmes (5.6.7) et (5.6.8), nous avons

$$\delta_{\mathbf{x}}\delta_{\mathbf{z}} \leq \frac{\kappa}{2}\left(\frac{\gamma^2 + (1-\sigma)^2 2r}{1-\gamma}\right)\mu$$

et par le lemme (5.6.5),

$$\begin{aligned}\alpha^* &\geq \frac{\gamma\sigma\mu}{2\delta_x\delta_z} \geq \frac{\gamma\sigma\mu 2(1-\gamma)}{2(\gamma^2+(1-\sigma)^2 2r)\kappa\mu} = \frac{\gamma\sigma(1-\gamma)}{(\gamma^2+(1-\sigma)^2 2r)\kappa} \\ &= \frac{\gamma(1-\delta/\sqrt{r})(1-\gamma)}{(\gamma^2+2\delta^2)\kappa} \geq \frac{\gamma(1-\gamma)}{c(\gamma^2+2\delta^2)\kappa}.\end{aligned}$$

où le terme $\sigma = 1-\delta/\sqrt{r}$ peut être borné inférieurement par une fraction $1/c > 0$ puisque par hypothèse $\sigma > 0$. Donc, puisque

$$\mu(\alpha) = (1-\alpha(1-\sigma))\mu = (1-\alpha\delta/\sqrt{r})\mu,$$

la réduction de μ à chaque itération vaut $1-\mathcal{O}(1/\kappa\sqrt{r})$. Nous avons alors,

$$\mu^k \leq \left(1-\frac{1}{\kappa\sqrt{r}}\right)\mu^{k+1} \leq \ldots \leq \left(1-\frac{1}{\kappa\sqrt{r}}\right)^k \mu^0.$$

D'où, $\mu^k \leq \epsilon\mu^0$ si $\left(1-\frac{1}{\kappa\sqrt{r}}\right)^k \leq \epsilon$, ou encore en prenant le logarithme, si

$$k\log\left(1-\frac{1}{\kappa\sqrt{r}}\right) \leq \log\epsilon.$$

De plus, le nombre de conditionnement κ étant toujours plus grand que 1, nous avons $1/\kappa\sqrt{r} < 1$ et donc $1-1/\kappa\sqrt{r} > 0$; nous pouvons alors majorer le terme $\log\left(1-\frac{1}{\kappa\sqrt{r}}\right)$ par $-1/\kappa\sqrt{r}$.
Par conséquent, $\mu^k \leq \epsilon\mu^0$ si $k(-1/\kappa\sqrt{r}) \leq \log\epsilon = -|\log\epsilon|$ c'est-à-dire si

$$k \geq \kappa\sqrt{r}|\log\epsilon| \geq \kappa\sqrt{r}\log\epsilon^{-1}.$$

Pour les algorithmes à pas semi-longs et à longs pas (i.e., pour les voisinages $\mathcal{N}_2(\gamma)$ et $\mathcal{N}_{-\infty}(\gamma)$) nous avons

$$\delta_x\delta_z \leq \frac{\kappa}{2}\left(1-2\sigma+\frac{\sigma^2}{1-\gamma}\right)2r\mu.$$

et

$$\alpha^* = \frac{\gamma\sigma\mu}{\delta_x\delta_z} \geq \frac{2\gamma\sigma\mu}{2\kappa\left(1-2\sigma+\frac{\sigma^2}{1-\gamma}\right)r\mu} = \frac{\gamma\sigma}{\kappa r\left(1-2\sigma+\frac{\sigma^2}{1-\gamma}\right)},$$

où σ est indépendant de r. L'inégalité obtenue montre que la réduction de μ à chaque itération vaut $1-\mathcal{O}(1/\kappa r)$. Ainsi, nous établissons de façon tout-à-fait similaire au cas précédent que ces deux algorithmes se terminent en $\mathcal{O}(\kappa r\log\epsilon^{-1})$ itérations.

□

Corollaire 5.6.1 *Si dans les algorithmes de suivi de chemins,* $\mathbf{p} \in \mathcal{C}(\mathbf{x}, \mathbf{z})$ *est choisi de sorte que* $\widetilde{\mathbf{x}} = \underset{\sim}{\mathbf{z}}$ *(méthode de Nesterov-Todd), alors les complexités algorithmiques sont les suivantes :*

1. $\mathcal{O}(\sqrt{r}\log\epsilon^{-1})$, *pour l'algorithme à petit pas.*
2. $\mathcal{O}(r\log\epsilon^{-1})$, *pour les algorithmes à pas semi-longs et à longs pas.*

Preuve : Ces deux résultats sont des conséquences immédiates du lemme (5.6.9) et du théorème (5.6.1)

□

Corollaire 5.6.2 *Si dans les algorithmes de suivi de chemins,* $\mathbf{p} \in \mathcal{C}(\mathbf{x}, \mathbf{z})$ *est choisi de sorte que* $\underset{\sim}{\mathbf{z}} = \mathbf{e}$ *(méthode* $\mathbf{xz}$*) ou* $\widetilde{\mathbf{x}} = \mathbf{e}$ *(méthode* $\mathbf{zx}$*), alors les complexités algorithmiques sont les suivantes :*

1. $\mathcal{O}(\sqrt{r}\log\epsilon^{-1})$, *pour l'algorithme à petit pas.*
2. $\mathcal{O}(r\log\epsilon^{-1})$, *pour l'algorithme à pas semi-longs.*
3. $\mathcal{O}(r^{1.5}\log\epsilon^{-1})$, *pour l'algorithme à longs pas.*

Preuve : Les résultats sont des conséquences directes du théorème (5.6.1) et des bornes sur $cond(G)$ établies par le lemme (5.6.9).

□

Conclusion

Ce mémoire a aboutit à l'élaboration d'algorithmes de suivi de chemins à petits pas, pas semi-longs et longs, tous de complexité générale polynomiale, pour la résolution de problèmes SOCP. Ceci a été réalisé par un passage essentiel par l'étude de l'algèbre de Jordan du cône du second ordre. Plusieurs extensions et travaux de recherche peuvent être réalisés afin d'éclairer certains points qui n'ont pas été abordés dans ce mémoire. Tout d'abord, il serait intéressant de rechercher des méthodes efficaces et numériquement stables pour la résolution du système linéaire (5.7), qui constitue la partie la plus coûteuse d'un point de vue numérique dans nos algorithmes. A ce sujet, certaines investigations (par exemple, [12]) suggèrent des méthodes basées sur la factorisation de Cholesky de la matrice AFA^T où $F = Q_{\mathbf{p}^{-1}}(Arw^{-1}(\mathbf{z})Arw(\tilde{\mathbf{x}}))Q_{\mathbf{p}^{-1}}$ et exploitant le caractère creux ou dense de A.

Une autre extension possible consisterait à étudier la façon dont pourrait se généraliser au cas de la programmation du cône du second ordre les algorithmes prédicteur-correcteur, de points intérieurs non-admissibles et de Mehrotra, bien connus en programmation linéaire. La connaissance de tels algorithmes ne ferait que confirmer que la programmation du cône du second ordre constitue une généralisation de la programmation linéaire.

Enfin, une question intéressante est de se demander si il est possible de développer une méthode du type simplexe pour la programmation du cône du second ordre. Un avantage que possèdent les algorithmes de type simplexe sur les méthodes de points intérieurs en programmation linéaire est que dès qu'un problème est résolu, et par la suite légèrement modifié, la méthode du simplexe duale peut être utilisée pour trouver la solution optimale à partir de la solution du problème d'origine. Typiquement, un nombre relativement faible d'itérations duales du simplexe suffisent, et chacune de ces itérations requiert $\mathcal{O}(m^2)$ opérations arithmétiques pour des problèmes denses et moins pour des problèmes creux. Pour les méthodes de points intérieurs, chaque itération coûte typiquement $\mathcal{O}(m^3)$ pour des problèmes denses ; d'où, même une seule itération s'avère coûteuse. Actuellement, aucune technique satisfaisante qui puisse utiliser la connaissance de la solution d'un problème légèrement différent et requièrant uniquement $\mathcal{O}(m^2)$ itérations arithmétiques n'est connue. En outre, aucun algorithme du type simplexe pour la programmation du cône du second ordre n'a

été proposé, de même qu'aucune des méthodes d'ensembles actifs pour la programmation non-linéaire n'a été explorée pour cette même classe de problèmes.

Annexes

Annexe A

Preuve de la $\mathcal{S}$-procédure

A.1 Lemmes préparatoires

Lemme A.1.1 *Un sous-ensemble C de $\mathbb{R}^n$ est un cône convexe si et seulement si $\forall x, y \in C$ et $\forall a, b > 0$, $ax + by \in C$.*

Preuve :
$\Leftarrow$: Soient $x \in C$ et $\alpha > 0$. En choisissant $a = b = \frac{\alpha}{2} > 0$ et $y = x$ nous avons que $\alpha x = \frac{\alpha}{2}x + \frac{\alpha}{2}x \in C$, d'où C est un cône. Le caractère convexe de C est évident.

$\Rightarrow$: Soient $x, y \in C$ et $a, b > 0$. Le résultat s'obtient immédiatement en se servant du fait que C est un cône convexe et en écrivant

$$ax + by = (1+a)\left[\left(\frac{a}{1+a}\right)x + \left(\frac{1}{1+a}\right)by\right] \in C.$$

□

Lemme A.1.2 *Soient $P(x)$ et $Q(x)$ deux formes quadratiques sur $\mathbb{R}^n$, i.e., $P(x) = x^T Ax$ et $Q(x) = x^T Bx$ pour des matrices symétriques données A, B et $\forall x \in \mathbb{R}^n$. Alors, l'ensemble*

$$W = \{(P(x);\, Q(x)) \in \mathbb{R}^2 \mid x \in \mathbb{R}^n\}$$

est un cône convexe.

Preuve : 1)W est un cône : Soit $(u;\, v) = (P(x);\, Q(x)) \in W$ pour un certain $x \in \mathbb{R}^n$. Pour tout $\alpha \geq 0$ nous avons

$$\alpha(u;\, v) = (\alpha P(x);\, \alpha Q(x)) = (P(\sqrt{\alpha}x);\, Q(\sqrt{\alpha}x)) \in W,$$

ce qui montre que W est un cône.

2)W est un cône convexe : En vertu du lemme (A.1.1), nous devons montrer que $aw' + bw'' \in W$ pour $w' = (u'; v')$ et $w'' = (u''; v'')$ dans W et $a, b > 0$, tous arbitraires.

- Cas où w' et w'' sont linéairement dépendants.
Nous avons, $w' = \alpha w''$ pour un certain $\alpha \in \mathbb{R}$. Deux cas peuvent se produire : soit $a\alpha + b \geq 0$, soit $a\alpha + b < 0$. Dans le premier cas, nous avons :

$$aw' + bw'' = a\alpha w'' + bw'' = (a\alpha + b)w'' \in W$$

car W est un cône. Dans l'autre cas ($a\alpha + b < 0$), nous avons nécessairement $\alpha < 0$ et donc

$$aw' + bw'' = (a\alpha + b)w'' = (a\alpha + b)\frac{1}{\alpha}w' \in W.$$

- Cas où w' et w'' sont linéairement indépendants.
Introduisons les formes quadratiques à deux variables $P(x, y)$ et $Q(x, y)$ définies par

$$P(x, y) = x^T Ay \quad \text{et} \quad Q(x, y) = x^T By \quad \forall\, x, y \in \mathbb{R}^n.$$

Puisque w' et w'' sont linéairement indépendants dans $\mathbb{R}^2$, tout vecteur de $\mathbb{R}^2$ peut s'écrire comme une combinaison linéaire unique de ces deux vecteurs. De plus, par définition de w' et w'', il existe x et $y \in \mathbb{R}^n$ tels que $w' = (u'; v') = (P(x); Q(x))$, $w'' = (u''; v'') = (P(y); Q(y))$. Avec x et y ainsi définis, nous pouvons écrire

$$\begin{pmatrix} P(x, y) \\ Q(x, y) \end{pmatrix} = \rho \begin{pmatrix} u' \\ v' \end{pmatrix} + \sigma \begin{pmatrix} u'' \\ v'' \end{pmatrix} \in \mathbb{R}^2$$

avec $\rho, \sigma \in \mathbb{R}$. Nous avons également pour $\alpha, \beta \in \mathbb{R}$

$$\begin{aligned} P(\alpha x + \beta y) &= (\alpha x + \beta y)^T A(\alpha x + \beta y) \\ &= \alpha^2 x^T Ax + 2\alpha\beta x^T Ay + \beta^2 y^T Ay \\ &= \alpha^2 P(x) + 2\alpha\beta P(x, y) + \beta^2 P(y) \\ &= \alpha^2 u' + 2\alpha\beta(\rho u' + \sigma u'') + \beta^2 u'' \\ &= (\alpha^2 + 2\alpha\beta\rho)u' + (\beta^2 + 2\alpha\beta\sigma)u'' \\ &= au' + bu'' \end{aligned}$$

sous les conditions

$$a = \alpha^2 + 2\alpha\beta\rho \tag{A.1}$$

$$b = \beta^2 + 2\alpha\beta\sigma \tag{A.2}$$

De façon similaire, nous obtenons

$$Q(\alpha x + \beta y) = av' + bv''$$

sous les mêmes conditions sur a et b.
Le résultat annoncé sera obtenu si nous parvenons à montrer que pour la paire (a, b) arbitraire, les équations (A.1) et (A.2) possèdent une solution (α, β) car alors nous aurons

$$\begin{aligned} aw' + bw'' &= a(u'; v') + b(u''; v'') \\ &= (au' + bu''; av' + bv'') \\ &= (P(\alpha x + \beta y); Q(\alpha x + \beta y)) \in W. \end{aligned}$$

Nous allons nous intéresser à une solution vérifiant $\beta = m\alpha$ avec $m \in \mathbb{R}$. Les équations (A.1) et (A.2) s'écrivent alors

$$a = \alpha^2(1 + 2m\rho) \quad \text{et} \quad b = \alpha^2(m^2 + 2m\sigma), \tag{A.3}$$

ce qui implique que $a(m^2 + 2m\sigma) = b(1 + 2m\rho)$ ou encore

$$am^2 + 2(a\sigma - b\rho)m - b = 0.$$

Remarquons que cette équation du second degré possède des racines m_1 et m_2 réelles puisque $\Delta = 4(a\sigma - b\rho)^2 + 4ab > 0$.

$$m_1 = \frac{(b\rho - a\sigma) + \sqrt{(a\sigma - b\rho)^2 + ab}}{a} > 0$$

$$m_2 = \frac{(b\rho - a\sigma) - \sqrt{(a\sigma - b\rho)^2 + ab}}{a} < 0.$$

Par conséquent, quel que soit le signe de ρ, il existe toujours une racine (m_1 ou m_2) telle que $m\rho \geq 0$ ce qui nous donne la solution (α, β) suivante établie à partir de (A.3) :

$$\alpha = \sqrt{\frac{a}{1 + 2m\rho}} \quad \text{et} \quad \beta = m\alpha.$$

□

Lemme A.1.3 *Soient C_1 et C_2 deux parties de $\mathbb{R}^n$ convexes, non vides et disjointes telles que C_1 soit un cône. Alors, il existe un hyperplan affine fermé séparant C_1 et C_2 au sens large, i.e., il existe un $z \in \mathbb{R}^n$ non nul tel que*

$$\begin{aligned} \langle x_1, z\rangle &\geq 0, \quad \forall x_1 \in C_1 \\ \langle x_2, z\rangle &\leq 0, \quad \forall x_2 \in C_2. \end{aligned}$$

Preuve : En appliquant le théorème de séparation des convexes au sens large (en dimension finie) à C_1 et C_2, nous obtenons

$$\exists z \in \mathbb{R}^n,\ z \neq 0 \text{ tq } \langle x_1, z\rangle \geq \langle x_2, z\rangle,\ \forall x_1 \in C_1 \text{ et } \forall x_2 \in C_2. \tag{A.4}$$

Soit $\alpha = \inf_{x_1 \in C_1} \langle x_1, z\rangle$. Puisque C_1 est un cône, nous avons

$$\lambda\langle x_1, z\rangle = \langle \lambda x_1, z\rangle \geq \alpha, \quad \forall x_1 \in C_1 \text{ et } \forall \lambda > 0.$$

En faisant tendre λ vers 0 par valeurs plus grandes, nous déduisons que $\alpha \leq 0$. De plus, par définition de l'infimum, si α était strictement négatif, nous aurions

$$\exists x_1 \in C_1 \text{ et } \exists \varepsilon > 0 \text{ tq } \langle x_1, z\rangle < -\varepsilon.$$

Ainsi, pour $\lambda > 0$ suffisamment grand,

$$\lambda\langle x_1, z\rangle = \langle \lambda x_1, z\rangle < \langle x_2, z\rangle \quad \text{pour un certain } x_2 \in C_2$$

ce qui est impossible en vertu de l'inégalité dans (A.4). D'où, $\alpha = 0$. Nous avons donc montré que $\langle x_1, z\rangle \geq 0, \forall x_1 \in C_1$ et $\langle x_2, z\rangle \leq 0, \forall x_2 \in C_2$.

□

A.2 Preuve de la $\mathcal{S}$-procédure

Proposition A.2.1 ($\mathcal{S}$-**procédure**)
Soient $F_0(x) = x^T Q_0 x + 2s_0^T x + r_0$ et $F_1(x) = x^T Q_1 x + 2s_1^T x + r_1$ deux fonctions quadratiques définies sur $\mathbb{R}^n$ et à valeurs dans $\mathbb{R}$ telles que Q_0 et Q_1 soient des matrices symétriques et telles qu'il existe $\tilde{x} \in \mathbb{R}^n$ vérifiant $F_1(\tilde{x}) > 0$. Alors, les deux affirmations suivantes sont équivalentes :

1. *$F_0(x) \geq 0 \quad \forall x \in \mathbb{R}^n$ tel que $F_1(x) \geq 0$.*
2. *Il existe $\tau \geq 0$ tel que*

$$\begin{bmatrix} Q_0 & s_0 \\ s_0^T & r_0 \end{bmatrix} - \tau \begin{bmatrix} Q_1 & s_1 \\ s_1^T & r_1 \end{bmatrix} \succcurlyeq 0.$$

Preuve : Nous commençons par donner une formulation équivalente de l'assertion 2.

$$2. \quad \Leftrightarrow \quad \exists \tau \geq 0 \text{ tq } y^T \begin{bmatrix} Q_0 - \tau Q_1 & s_0 - \tau s_1 \\ s_0^T - \tau s_1^T & r_0 - \tau r_1 \end{bmatrix} y \geq 0, \ \forall y \in I\!R^{n+1} \qquad (A.5)$$

$$\Leftrightarrow \quad \exists \tau \geq 0 \text{ tq } \begin{bmatrix} y \\ 1 \end{bmatrix}^T \begin{bmatrix} Q_0 - \tau Q_1 & s_0 - \tau s_1 \\ s_0^T - \tau s_1^T & r_0 - \tau r_1 \end{bmatrix} \begin{bmatrix} y \\ 1 \end{bmatrix} \geq 0, \ \forall y \in I\!R^n \qquad (A.6)$$

$$\Leftrightarrow \quad \exists \tau \geq 0 \text{ tq } F_0(y) - \tau F_1(y) \geq 0, \ \forall y \in I\!R^n. \qquad (A.7)$$

Notons que l'implication (A.6) $\Rightarrow$ (A.5) est vérifiée de façon évidente pour des vecteurs $y \in I\!R^{n+1}$ tels que $y_{n+1} \neq 0$ en normalisant cette dernière composante. Pour des vecteurs $y \in I\!R^{n+1}$ dont la composante $y_{n+1} = 0$ le résultat est également vérifié puisqu'il l'est pour les vecteurs $y \in I\!R^{n+1}$ du type $(y_1; \ldots; y_n; \varepsilon)$ avec $\varepsilon \neq 0$ et par continuité.

2 $\Rightarrow$ 1 : En vertu de la reformulation de l'assertion 2, l'implication dans ce sens est triviale.

1 $\Rightarrow$ 2 : L'implication dans ce sens est beaucoup moins triviale et demandera l'utilisation des lemmes précédemment établis.

A) L'implication est vraie pour deux formes quadratiques : Considérons deux formes quadratiques arbitraires $P(x) = x^T Ax$ et $Q(x) = x^T Bx$ définies sur $I\!R^n$, à valeurs réelles, vérifiant l'assertion 1 avec des matrices A et B symétriques. Grâce au lemme (A.1.2), nous avons que l'ensemble $W = \{(P(x); Q(x)) \in I\!R^2 \mid x \in I\!R^n\}$ est un cône convexe, non vide, disjoint (en vertu de l'assertion 1) du cône convexe $Q = \{(a; b) \in I\!R^2 \mid a < 0, b \geq 0\}$. Par conséquent, par le lemme (A.1.3), il existe $z = (z_1; -z_2) \in I\!R^2$, $z \neq 0$ tel que

$$z_1\eta_1 - z_2\eta_2 \geq 0 \quad \forall (\eta_1; \eta_2) \in W, \qquad (A.8)$$

$$z_1\eta_1 - z_2\eta_2 \leq 0 \quad \forall (\eta_1; \eta_2) \in Q. \qquad (A.9)$$

Puisque $(-1; 0) \in Q$ et $(-\varepsilon; 1) \in Q$, où $\varepsilon > 0$ est arbitrairement petit, il suit en utilisant (A.9) que $z_1 \geq 0$ et $z_2 \geq 0$. Puisque $\exists \tilde{x} \in I\!R^n$ tq $Q(\tilde{x}) > 0$, nous avons que $z_1 \neq 0$ (autrement, par (A.8), nous aurions $0 \geq -z_2 Q(\tilde{x}) \geq 0$ et donc $z_2 Q(\tilde{x}) = 0$, ce qui est impossible). Ainsi, nous obtenons $\eta_1 - \tau\eta_2 \geq 0$ pour tout $(\eta_1; \eta_2) \in W$ où $\tau = z_2/z_1 \geq 0$, d'où, l'assertion 2 est vérifiée.

B) L'implication est vraie pour deux fonctions quadratiques : Soient F_0 et F_1 les deux fonctions quadratiques vérifiant l'assertion 1. Nous pouvons, s.p.d.g., supposer que $\tilde{x} = 0$ (sinon, nous définissons $\widehat{F}_0(x) = F_0(x + \tilde{x})$ et $\widehat{F}_1(x) = F_1(x+\tilde{x})$ vérifiant $\widehat{F}_1(0) > 0$). Définissons à présent les deux formes quadratiques suivantes définies sur $I\!R^n \times I\!R$ et correspondant respectivement à F_0 et F_1 :

$$G_0(x, \zeta) = x^T Q_0 x + 2s_0^T x\zeta + r_0\zeta^2 = \begin{bmatrix} x \\ \zeta \end{bmatrix}^T \begin{bmatrix} Q_0 & s_0 \\ s_0^T & r_0 \end{bmatrix} \begin{bmatrix} x \\ \zeta \end{bmatrix},$$

$$G_1(x, \zeta) = x^T Q_1 x + 2 s_1^T x \zeta + r_1 \zeta^2 = \begin{bmatrix} x \\ \zeta \end{bmatrix}^T \begin{bmatrix} Q_1 & s_1 \\ s_1^T & r_1 \end{bmatrix} \begin{bmatrix} x \\ \zeta \end{bmatrix},$$

$\forall (x, \zeta) \in \mathbb{R}^n \times \mathbb{R}$. L'idée est de se ramener au cas particulier A) avec les formes quadratiques G_0 et G_1 ; nous allons montrer que $G_0(x, \zeta) \geq 0$ lorsque $G_1(x, \zeta) \geq 0$ où $G_1(\tilde{x}, 1) = G_1(0, 1) = F_1(0) > 0$.
Pour les couples (x, ζ) où $\zeta \neq 0$, cela est une conséquence de l'assertion 1, puisque

$$G_0(x, \zeta) = \zeta^2 F_0(\zeta^{-1} x) \text{ et } G_1(x, \zeta) = \zeta^2 F_1(\zeta^{-1} x).$$

Soit $G_1(x, 0) \geq 0$ (*i.e.*, $x^T Q_1 x \geq 0$) pour un certain x. Puisque $F_1(\tilde{x}) = F_1(0) = r_1 > 0$, il suit que $F_1(\eta x) = \eta^2 x^T Q_1 x + 2\eta b_1^T x + r_1 \geq 0$ pour des valeurs de $\eta \in \mathbb{R}$ suffisamment grandes en valeur absolue (si $x^T Q_1 x > 0$, cela est évident ; si $x^T Q_1 x = 0$ et $b_1^T x = 0$, nous avons $F_1(\eta x) = r_1 > 0$ pour tout η ; si $x^T Q_1 x = 0$ et $b_1^T x \neq 0$, alors η doit être choisi de signe adéquat et suffisamment grand). Ainsi, par l'assertion 1 nous avons $F_0(\eta x) \geq 0$ et donc $G_0(x, 0) = x^T Q_0 x = \lim_{|\eta| \to +\infty} \eta^{-2} F_0(\eta x) \geq 0$ avec η de signe adéquatement choisi. Par conséquent, en utilisant la partie A) avec $P := G_0$ et $Q := G_1$, nous avons

$$\exists \tau \geq 0 \text{ tq } G_0(x, \zeta) - \tau G_1(x, \zeta) \geq 0, \quad \forall (x, \zeta) \in \mathbb{R}^n \times \mathbb{R}.$$

L'assertion 2 est obtenue avec $\zeta = 1$.

□

Annexe B

Applications

B.0.1 Synthèse d'un tableau d'antennes

La synthèse d'un tableau d'antennes fait partie des nombreux problèmes d'ingénierie pouvant être modélisés par un problème SOCP. Une antenne est un appareil électromagnétique capable d'émettre ou de recevoir des ondes électromagnétiques. La caractéristique principale d'une antenne monochromatique est son diagramme $Z(\delta)$, qui est une fonction à valeurs complexes d'une direction à trois dimension δ. La densité directionnelle de l'énergie emmagasinée par l'onde emise dans la direction δ est fonction de la valeur absolue $|Z(\delta)|$ propre au diagramme ; elle est en fait proportionnelle à $|Z(\delta)|^2$. L'argument $\arg Z(\delta)$ du diagramme correspond à la phase initiale de l'onde se propageant dans la direction δ, de sorte que le champ électrique généré par l'antenne, perçu au temps t à un point $P = r\delta$ (nous supposons que l'antenne est placée à l'origine) est proportionnel à

$$E(r\delta, t) = |Z(\delta)|r^{-1}\cos(\arg Z(\delta) + t\omega - 2\pi r/\lambda), \tag{B.1}$$

où ω désigne la fréquence de l'onde et λ, la longueur d'onde.[1] Pour une antenne complexe constituée de n antennes de diagrammes $Z_1(\delta), \ldots, Z_n(\delta)$, le diagramme résultant $Z(\cdot)$ est par définition

$$Z(\delta) = \sum_{j=1}^{n} Z_j(\delta).$$

Lorsqu'un ingénieur conçoit un tableau d'antennes comportant plusieurs antennes, il commence avec n "blocs constructeurs" de diagrammes $Z_1, \ldots, Z_n$. Pour chacun d'eux, l'ingénieur peut augmenter l'amplitude $a_j(\cdot)$ du signal envoyé d'un facteur ρ_j et décaler la phase initiale $\phi_j(\cdot)$ d'un facteur constant ψ_j. En d'autres mots, l'ingénieur peut modifier les diagrammes d'origine des blocs

[1] La relation (B.1) est valable lorsque la distance entre P et l'antenne est beaucoup plus grande que la dimension de l'antenne. La différence entre le membre de gauche et le membre de droite dans (B.1) est $o(r^{-1})$ lorsque $r \to \infty$.

selon

$$Z_j(\delta) \equiv a_j(\delta)[\cos(\phi_j(\delta)) + i\sin(\phi_j(\delta))]$$
$$\mapsto Z_j^+(\delta) = \rho_j a_j(\delta)[\cos(\phi_j(\delta) + \psi_j) + i\sin(\phi_j(\delta) + \psi_j)].$$

Donc, il est possible de multiplier le diagramme initial de chaque bloc par une constante arbitraire complexe

$$z_j = \rho_j(\cos\psi_j + i\sin\psi_j) \equiv u_j + iv_j.$$

Le diagramme de l'antenne complexe résultante sera

$$Z(\delta) = \sum_{j=1}^{n} z_j Z_j(\delta). \tag{B.2}$$

Le problème de la synthèse d'un tableau d'antennes associé à la description donnée est de choisir adéquatement les paramètres z_j, $j = 1,\ldots,n$, de sorte que l'on obtienne un diagramme aussi proche que possible d'un diagramme "cible" $Z_*(\delta)$. Ce genre de problème se rencontre, par exemple, dans la marine où le choix d'une direction "d'écoute" donnée exige de se régler sur un signal adéquat et de supprimer tout autre signal provenant d'autres directions . Dans beaucoup de cas, une mesure pertinente est faite à l'aide de la norme uniforme, et le problème de synthèse s'exprime alors comme

$$\min_{z_1,\ldots,z_n\in\mathbb{C}^n} \max_{\delta:\|\delta\|_2=1} |Z_*(\delta) - \sum_{j=1}^{n} z_j Z_j(\delta)|,$$

où les variables de modélisation sont les n nombres complexes $z_1,\ldots,z_n$, ou, ce qui revient au même, les $2n$ nombres réels $Re(z_j)$, $Im(z_j)$. A partir de la figure B.1, nous pouvons observer dans (a) 10 éléments du tableau d'antennes sous forme d'anneaux dans le plan xy; dans (b) sont représentés les "blocs de construction"-les diagrammes des anneaux comme des fonctions d'un angle d'altitude δ. Dans la figure B.2, deux diagrammes sont représentés; celui en pointillés correspond au diagramme cible et celui en trait plein est le diagramme synthétisé.

Plaçons-nous dans le cadre particulier où le tableau est situé dans un plan et constitué de 10 antennes similaires à celles de la figure (a). Cette hypothèse permet d'établir, grâce à des lois physiques, que le diagramme résultant de ces antennes est une fonction à valeurs réelles dépendant uniquement de l'angle d'altitude θ entre le plan et la direction δ. Aussi, pour pouvoir aboutir à un problème SOCP, il est nécessaire de supposer une discrétisation de cet angle θ en $\theta_1,\ldots,\theta_N$ où $N \gg n$. Ces hypothèses mènent au modèle suivant à 11 variables réelles appelé problème nominal :

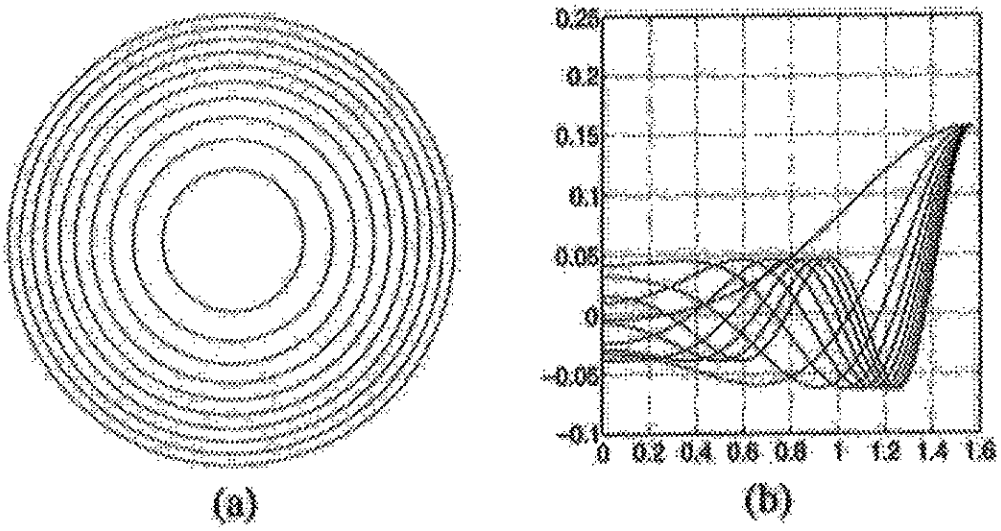


figure B.1: Exemple d'antennes (a) et de "blocs constructeurs" (b)

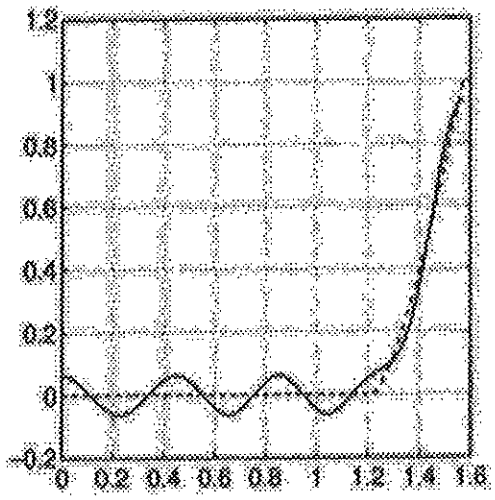


figure B.2: Diagramme synthétisé (en trait plein) et diagramme "cible" (en traits pointillés)

$$\min_{x,t}\left\{t\,|\,-t \le Z_*(\theta_i) - \sum_{j=1}^{10} x_j Z_j(\theta_i) \le t,\ i=1,\ldots,N\right\}. \qquad \text{(B.3)}$$

Désignons par x_j^* les valeurs optimales des variables de modélisation. Rappelons qu'il s'agit de coefficients d'amplification ce qui signifie qu'ils dépendent des caractériques de certains appareils physiques. Dans la réalité, bien sûr, nous ne savons pas régler les appareils aux valeurs précises x_j^* ; le mieux que nous puissions espérer est que les valeurs x_j^r que l'on peut obtenir réellement coïncident avec les valeurs désirées x_j^* avec une marge d'erreur, par exemple, de 0.1% :

$$x_j^r = p_j x_j^*, \quad 0.999 \le p_j \le 1.001, \quad j = 1,\ldots,10,$$

où les p_j sont supposés être des nombres aléatoires indépendants de moyenne 1. Il est évident que du fait du caractère aléatoire des p_j et donc des x_j^r, le diagramme (réaliste) obtenu avec ces coefficients diffèrera du diagramme (idéaliste) obtenu avec les valeurs optimales x_j^*. De façon très surprenante, il s'avère que même en utilisant un échantillon de 10 nombres aléatoires p_j distribués uniformément sur [0.999, 0.001] nous obtenons un diagramme $Z_r(\theta_i) = \sum_{j=1}^{10} p_j x_j^* Z_j(\theta_i)$ extrêmement différent (voir figure B.3 graphique de droite - tracé en trait plein) du diagramme $Z_*(\theta_i) = \sum_{j=1}^{10} x_j^* Z(\theta_i)$ (voir figure B.3 graphique de gauche - tracé en trait plein). Nous voyons que la différence entre les deux diagrammes est tant au niveau de la forme générale de la courbe qu'au niveau de l'étendue des valeurs.

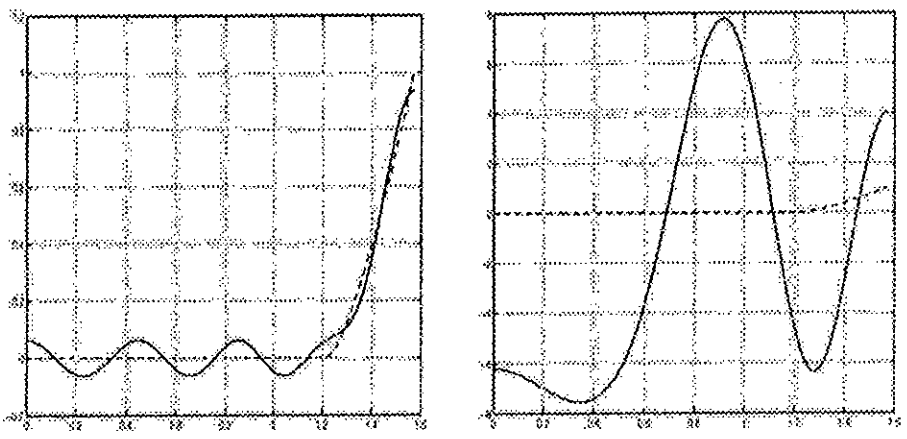


figure B.3: A gauche, le diagramme idéal correspondant aux coefficients x_j^* (en trait plein) ; à droite, le diagramme réaliste correspondant aux coefficients x_j^r (en trait plein). Dans les deux figures, le diagramme en pointillés correspond au diagramme "cible" duquel il faut se rapprocher le plus possible.

Afin d'obtenir de meilleurs résultats, essayons d'appliquer une approche robuste à notre problème. Pour cela, nous supposerons dès le début que les valeurs à atteindre sont les coefficients x_j, alors que les coefficients réels (ou réalistes) d'amplification seront les $x_j^r = p_j x_j$, $0.999 \leq p_j \leq 1.001$, $j = 1, \ldots, 10$ et les écarts des diagrammes pour chaque angle θ_i, $i = 1, \ldots, N$ seront donnés par

$$\delta_i(x) = Z_*(\theta_i) - \sum_{j=1}^{10} p_j x_j Z_j(\theta_i),$$

où $x \equiv (x_1, \ldots, x_{10})$. Nous sommes en fait devant un problème de programmation linéaire robuste, où l'incertitude se situe au niveau des coefficients d'amplification. Ces coefficients varient avec une marge d'erreur de 0.1% par rapport à leurs valeurs idéales.
Pour utiliser l'indépendance statistique des perturbations, voyons ce qu'il advient d'une inégalité de la forme

$$-t \leq \delta_i(x) \equiv Z_*(\theta_i) - \sum_{j=1}^{10} p_j x_j Z_j(\theta_i) \leq t, \tag{B.4}$$

lorsque les p_j sont aléatoires. Pour un x fixé, l'écart $\delta_i(x)$ est une variable aléatoire d'espérance

$$\delta_i^*(x) = Z_*(\theta_i) - \sum_{j=1}^{10} x_j Z_j(\theta_i)$$

et d'écart type

$$\sigma_i(x) = \sqrt{E\{(\delta_i(x) - \delta_i^*(x))^2\}}.$$

L'espérance $E\{(\delta_i(x) - \delta_i^*(x))^2\}$ peut être développée plus en détail en se servant de l'indépendance statistique des p_j et donc, en particulier, du fait que la covariance $\mathrm{Cov}(p_k, p_j) = 0$, $\forall k \neq j$. Nous avons alors

$$\begin{aligned}
E\{(\delta_i(x) - \delta_i^*(x))^2\} &= E\left\{\left(\sum_{j=1}^{10} x_j Z_j(\theta_i)(p_j - 1)\right)^2\right\} \\
&= E\left\{\left(\sum_{j=1}^{10} x_j^2 Z_j^2(\theta_i)(p_j - 1)^2 + 2\sum_{k<j} x_k Z_k(\theta_i)(p_k - 1) x_j Z_j(\theta_i)(p_j - 1)\right)\right\} \\
&= \sum_{j=1}^{10} x_j^2 Z_j^2(\theta_i) E\{(p_j - 1)^2\} + 2\sum_{k<j} x_k Z_k(\theta_i) x_j Z_j(\theta_i) \underbrace{\mathrm{Cov}(p_k, p_j)}_{=0} \\
&= \sum_{j=1}^{10} x_j^2 Z_j^2(\theta_i) E\{(p_j - 1)^2\}
\end{aligned}$$

en notant que l'espérance de $(p_j - 1)^2$ peut être majoré de la sorte

$$\begin{aligned} E\{(p_j-1)^2\} &= \int_{0.999}^{1.001}(p_j-1)^2 f_{p_j}\,dp_j \\ &\leq \int_{0.999}^{1.001}(p_j-1)^2\,dp_j \\ &= \left[\frac{(p_j-1)^3}{3}\right]_{0.999}^{1.001} = \frac{2}{3}(0.001)^3 \approx 6.66667.10^{-10} \end{aligned}$$

où f_{p_j} désigne la fonction de densité de la variable aléatoire p_j dont la loi de probabilité est supposée inconnue. Comme cette densité représente une probabilité, elle peut être majorée par 1 dans l'intégrale. Nous obtenons par conséquent,

$$\sigma_i(x) = \sqrt{\sum_{j=1}^{10} x_j^2 Z_j^2(\theta_i) E\{(p_j-1)^2\}} \leq \sqrt{6.66667.10^{-10}}\sqrt{\sum_{j=1}^{10} x_j^2 Z_j^2(\theta_i)} \leq \kappa\nu_i(x),$$

$$\text{avec } \nu_i(x) = \sqrt{\sum_{j=1}^{10} x_j^2 Z_j^2(\theta_i)} \text{ et } \kappa = 0.001.$$

Une valeur typique de $\delta_i(x)$ diffèrera de $\delta_i^*(x)$ d'une quantité de l'ordre de $\sigma_i(x)$. Nous allons ensuite fixer un paramètre de sécurité $\omega > 00$ de sorte que tous les vecteurs x donnant lieu à $|\delta_i(x) - \delta_i^*(x)| > \omega\kappa\nu_i(x)$ seront ignorés.[2] Pour les x donnant lieu à $|\delta_i(x) - \delta_i^*(x)| \leq \omega\kappa\nu_i(x)$ nous allons modifier les inégalités (B.4) à la façon d'un ingénieur qui affirme qu'une variable aléatoire diffère, avec une très grande probabilité, de sa moyenne, d'au plus trois fois son écart type. Utilisant la borne $\kappa\nu_i$, nous obtenons

$$-t \leq \delta_i(x) \leq t \Leftrightarrow -t \leq \delta_i^*(x) \pm \omega\kappa\nu_i(x) \leq t$$

$$\Leftrightarrow \begin{cases} -t \leq \delta_i^*(x) - \omega\kappa\nu_i(x), \\ \delta_i^*(x) + \omega\kappa\nu_i(x) \leq t. \end{cases} \tag{B.5}$$

Remplaçant les inégalités du problème (B.3) par les inégalités (B.5) et posant $Q_i = \omega\kappa.diag(Z_1(\theta_i),\ldots,Z_{10}(\theta_i))$, $i = 1,\ldots,N$, nous aboutissons finalement au problème SOCP robuste suivant

$$\begin{cases} \min & t \\ \text{s.c.} & \|Q_i x\| \leq \left[Z_*(\theta_i) - \sum_{j=1}^{10} x_j Z_j(\theta_i)\right] + t,\ i = 1,\ldots,N, \\ & \|Q_i x\| \leq -\left[Z_*(\theta_i) - \sum_{j=1}^{10} x_j Z_j(\theta_i)\right] + t,\ i = 1,\ldots,N. \end{cases}$$

L'idée de reformuler le problème dans une version robuste était judicieuse puisque les diagrammes "aléatoires" obtenus à partir des solutions x_j^* du problème robuste sont maintenant très proche du diagramme cible ; en effet, sur un ensemble

[2] Il serait préférable d'utiliser l'écart type σ_i au lieu de ν_i. Cependant, ne connaissant pas la distribution des p_j, il est plus commode d'un point de vue pratique de remplacer les inconnues σ_i par leur borne supérieure ν_i.

de 40 diagrammes engendrés à partir des solutions robustes, les distances uniformes par rapport au diagramme cible s'étendent de 0.0814 à 0.0830, ce qui est incomparable avec les distances uniformes obtenues par la première approche ; sur le diagramme de la figure B.3, nous pouvons observer que cette distance atteind la valeur 7.7. La figure B.4 illustre un diagramme typique obtenu par l'approche robuste.

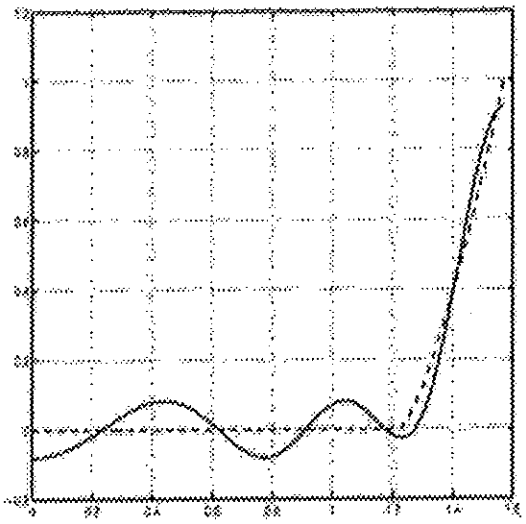


figure B.4: Diagramme robuste. La distance uniforme au diagramme cible est de 0.0822. Le paramètre de sécurité ω vaut 1.

Une autre constatation intéressante est que lorsque le pourcentage d'erreurs est modifié, disons qu'il augmente à 1%, la solution n'est pas fortement perturbée. En effet, sur un échantillon de 40 diagrammes obtenus par les anciennes solutions optimales robustes (désormais affectées d'erreurs d'implémentation à 1%), les distances uniformes au diagramme cible varient de 0.0834 à 0.116, ce qui est encore largement inférieur à la valeur 7.7.
Une question s'impose : pourquoi la solution du problème nominal (B.3) est-elle si instable et pourquoi l'approche robuste donne lieu à des résultats bien meilleurs ?
La réponse à cette question devient claire lorsque nous inspectons les valeurs optimales des coefficients d'amplification des problèmes nominal et robuste.

j	1	2	3	4	5	6
x_j^{nom}	1624.4	−14701	55383	−107247	95468	19221
x_j^{rob}	−0.3010	4.9638	−3.4252	−5.1488	6.8653	5.5140

j	7	8	9	10
x_j^{nom}	−138622	144870	−69303	13311
x_j^{rob}	5.3119	−7.4584	−8.9140	13.237

Puisque les imprécisions obtenues au moment de choisir les coefficients d'amplification (le plus proche possible de leurs valeurs optimales) sont d'autant plus grandes que ces valeurs optimales sont grandes, il n'est plus surprenant, au vu des valeurs des x_j^{nom} caractérisées par de grandes magnitudes, d'avoir obtenu de si mauvais résultats pour le problème nominal. Par contre, l'équivalent robuste pénalise les magnitudes des solutions (grâce aux termes $\|Q_i x\|$ qui figurent dans les contraintes du problème robuste) et donc, mène à une modélisation beaucoup plus stable.

B.0.2 Conception d'une topologie en trellis

Un trellis est une construction mécanique constituée de fines barres élastiques liées les unes aux autres, qui a un aspect similaire à un poteau électrique, un pont de chemin de fer, ou à la tour Eiffel. Les points de connexion entre les barres élastiques sont appelés des noeuds. Un trellis peut être soumis à une charge extérieure - un ensemble de forces agissant simultanément au niveau des noeuds, comme le montre la figure B.5

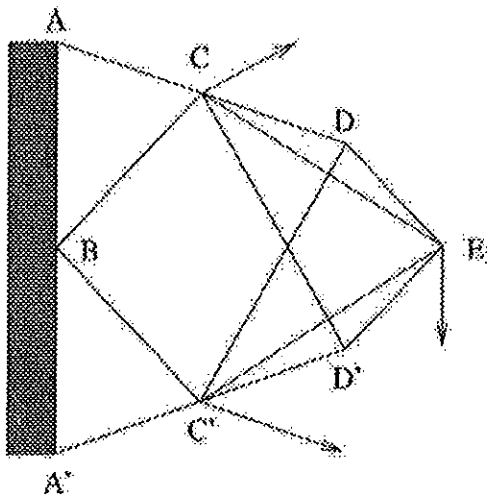


figure B.5: Un trellis à deux dimensions soumis à une charge extérieure. Les noeuds : A, A', B, C, C', D, D', E ; les barres : AC, A'C', BC, BC', CD, CD', C'D', C'D, CE, C'E, DE, D'E ; les forces : flèches.

Du fait de la présence d'une charge, la construction se déforme quelque peu, jusqu'à ce que les forces de tension dans les barres compensent exactement les forces extérieures. Etant déformé, le trellis emmagasine une certaine quantité d'énergie potentielle élastique qui sera d'autant plus petite que le trellis sera rigide.

Nous utiliserons désormais l'appellation "problème CTT" pour désigner le problème de conception d'une topologie en trellis. Les données pour la version de

base du problème CTT sont les suivantes :

- Un ensemble fini de noeuds (représentés par des points dans un espace à deux ou trois dimensions) ainsi qu'un ensemble fini de barres élastiques qui doivent relier les noeuds.
- Des conditions aux limites spécifiant les éventuels noeuds contraints à être fixes dans la structure (cfr noeuds A, B, A' dans la figure B.5).
- Une charge constituée d'un ensemble de forces extérieures agissant sur les noeuds.

L'objectif du problème CTT est de concevoir un trellis ne dépassant pas un poids donné qui supporte le mieux la charge, c'est-à-dire qu'il faut lier certaines paires de noeuds par des barres de tailles appropriées, ne dépassant pas un poids total donné, de sorte que l'énergie potentielle emmagasinée par le trellis soit aussi petite que possible.
Une propriété intéressante du problème CTT est qu'il détermine non seulement les poids des barres mais également la forme du trellis. En effet, supposons que nous ayons une grille nodale dense où toute paire de noeuds est reliée par une barre. Dans le trellis optimal, obtenu par le problème d'optimisation associé, certaines barres (typiquement, la majorité d'entre elles) se verront attribuer un poids nul. En d'autres mots, le problème d'optimisation décidera par lui-même quels noeuds utiliser et comment les lier, i.e., il trouvera à la fois le schéma optimal (topologie) de la construction ainsi que son poids optimal.

Etablissement du modèle

Pour poser le problème CTT comme un problème d'optimisation, voyons de façon plus précise ce qu'il advient d'un trellis lorsqu'il est soumis à une charge. Soit une barre quelconque AB dans le trellis non soumis à la charge (figure B.6). Une fois la charge appliquée, les noeuds A et B se déplacent légèrement dans les directions respectives dA et dB, comme le montre la figure B.6.

En supposant que les déplacements nodaux dA et dB soient petits et en négligeant les termes du second ordre, l'élongation dl de la barre due à la charge est la projection du vecteur $dB - dA$ sur la direction de la barre :

$$dl = (dB - dA)^T \frac{(B - A)}{\|B - A\|}.$$

La tension dans la barre AB causée par cette élongation est donnée grâce à la loi de Hooke, par :

$$\kappa \frac{dl \times S_{AB}}{\|B - A\|} = \kappa \frac{dl \times t_{AB}}{\|B - A\|^2},$$

où κ est le module de Young qui est une caractéristique du matériau, S_{AB} est l'aire de la section de la barre AB et t_{AB} est le volume de la barre AB. Ainsi, en se servant de l'expression de dl, la tension est

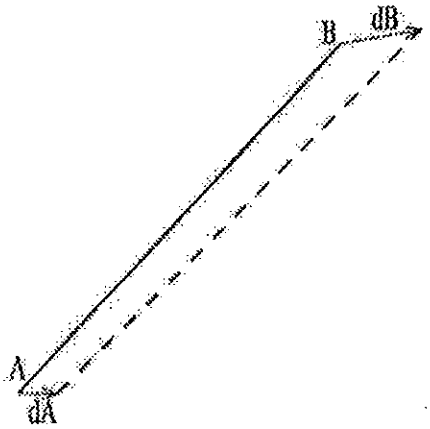


figure B.6: Une barre avant (trait plein) et après (traits pointillés) application d'une charge.

$$\tau = \kappa t_{AB}(dB - dA)^T(B - A)\|B - A\|^{-3}.$$

La force de réaction associée à la tension τ au point B est le vecteur

$$\begin{aligned} -\tau(B-A)\|B-A\|^{-1} &= -\kappa t_{AB}[(dB-dA)^T(B-A)](B-A)\|B-A\|^{-4} \\ &= -t_{AB}[(dB-dA)^T\beta_{AB}]\beta_{AB}, \end{aligned} \tag{B.6}$$

où $\beta_{AB} = \sqrt{\kappa}(B-A)\|B-A\|^{-2}$. Notons que le vecteur β_{AB} dépend des positions des noeuds liés par la barre et est indépendant de la charge présente ainsi que de la conception du trellis.
A présent, intéressons-nous à l'énergie potentielle emmagasinée par notre barre AB suite à l'élongation. Les physiciens ont établi que cette énergie vaut la moitié du produit entre la tension et l'élongation, i.e.,

$$\begin{aligned} E_{pot} = \frac{\tau dl}{2} &= \frac{[\kappa t_{AB}(dB-dA)^T(B-A)\|B-A\|^{-3}][(dB-dA)^T(B-A)\|B-A\|^{-1}]}{2} \\ &= \tfrac{1}{2}t_{AB}\left[(dB-dA)^T\beta_{AB}\right]^2. \end{aligned} \tag{B.7}$$

Désignons par M le nombre total de noeuds et par M_f le nombre de noeuds libres, c'est-à-dire ceux n'étant pas fixés par les conditions aux limites. Définissons l'espace $\mathbb{R}^m$ des *déplacements virtuels* (propre à la construction) comme étant le produit cartésien des espaces de déplacement des noeuds libres ; donc, m vaut soit $2M_f$, soit $3M_f$, suivant que nous considérions un trellis à deux ou trois dimensions. Un vecteur v de $\mathbb{R}^m$ représente un déplacement de la grille nodale : un noeud libre ν correspond à une paire (cas d'un trellis dans le plan) ou un triplet (cas d'un trellis tri-dimensionnel) de coordonnées de v, et le sous-vecteur correspondant $v[\nu]$ de v représente le déplacement "physique" (en deux ou trois

dimensions) du noeud libre ν. Il est utile de définir les sous-vecteurs $v[\nu]$ pour les noeuds fixés ; par définition, ces sous-vecteurs sont des vecteurs nuls.
Une charge - un ensemble de forces extérieures agissant sur les noeuds libres - peut être représentée par un vecteur $f \in \mathbb{R}^m$; pour chaque noeud libre ν, le sous-vecteur correspondant $f[\nu]$ de f est la force extérieure agissant sur ν.
Soit n le nombre de barres libres dans le trellis, i.e., les barres qui lient deux noeuds tels qu'au moins l'un des deux soit libre. Ordonnons ces n barres libres et considérons la i-ème d'entre elles. Cette barre lie deux noeuds $\nu'(i)$, $\nu''(i)$, i.e., deux points A_i et B_i de notre espace physique. Définissons à présent pour chaque barre libre i, un vecteur $b_i \in \mathbb{R}^m$ de la façon suivante (cfr. définition de β_{AB}) :

$$b_i[\nu] = \begin{cases} \beta_{A_iB_i}, & \nu = \nu''(i) \text{ et } \nu \text{ est libre,} \\ -\beta_{A_iB_i}, & \nu = \nu'(i) \text{ et } \nu \text{ est libre,} \\ 0 & \text{dans tous les autres cas.} \end{cases} \tag{B.8}$$

Un trellis particulier peut être identifié par un vecteur non-négatif $t = (t_1, \ldots, t_n)$, où t_i est le volume de la barre i dans le trellis. Ainsi, considérons un trellis t, et intéressons-nous aux forces de réaction causées par un déplacement v des noeuds du trellis. D'après (B.6) et (B.8), il suit que pour chaque noeud libre ν, la composante de la force de réaction causée, sous le déplacement, par la barre i sur le noeud ν est $-t_i(b_i^T v)b_i[\nu]$. Par conséquent, la force de réaction totale au niveau du noeud ν est

$$-\sum_{i=1}^{n} t_i(b_i^T v)b_i[\nu].$$

Notons que toute barre qui n'est pas incidente au noeud ν aura un terme correspondant qui est nul dans la somme en vertu de (B.8). L'ensemble des forces de réaction au niveau des noeuds est donc

$$-\sum_{i=1}^{n} t_i(b_i^T v)b_i = -\left[\sum_{i=1}^{n} t_i(b_i b_i^T)\right] v \in \mathbb{R}^m.$$

Ce vecteur de dimension m dépend linéairement du déplacement v :

$$f_r = -A(t)v$$

où

$$A(t) = \sum_{i=1}^{n} t_i b_i b_i^T$$

est appelée *matrice de raideur* du trellis. Il s'agit d'une matrice $m \times m$ symétrique, qui dépend linéairement des variables du problème t_i - les volumes des barres libres. Elle est également semi-définie positive puisque :

$$\forall\, d \in I\!R^m, \quad d^T \left(\sum_{i=1}^{n} t_i b_i b_i^T\right) d = \sum_{i=1}^{n} t_i (d^T b_i)(b_i^T d) = \sum_{i=1}^{n} t_i (b_i^T d)^2 \geq 0.$$

Rappelons qu'en vertu de la loi de l'action et la réaction, les forces de réaction doivent compenser les forces extérieures, c'est-à-dire : $f + f_r = 0$, ce qui nous donne un système d'équations linéaires qui détermine le déplacement du trellis dû à la charge f :

$$A(t)v = f. \tag{B.9}$$

Afin d'aboutir à un problème SOCP, nous faisons les deux hypothèses suivantes :

- La somme des matrices $b_i b_i^T$ est définie positive.
- Les t_i sont tous strictement positifs. Cepedant, afin reconnaître les barres i qui ne doivent pas être présentes dans le trellis, nous interpréterons un t_i strictement positif et proche de 0 comme l'absence de la barre i dans le trellis.

Ces hypothèses nous permettent d'affirmer, comme nous l'avons montré dans le point d) de la section 2.1.4, que $A(t)$ est inversible.
Pour terminer la mise en place du problème, nous devons réécrire l'expression de l'énergie potentielle emmagasinée dans le trellis à l'équilibre. Grâce à (B.7) et (B.8), cette énergie s'exprime comme :

$$\begin{aligned}
\frac{1}{2}\sum_{i=1}^{n} t_i \left[(v[\nu''(i)] - v[\nu'(i)])^T \beta_{A_i B_i}\right]^2 &= \tfrac{1}{2}\sum_{i=1}^{n} t_i (v^T b_i)^2 \\
&= \tfrac{1}{2}\sum_{i=1}^{n} t_i (v^T b_i)(b_i^T v) \\
&= \tfrac{1}{2} v^T \left[\sum_{i=1}^{n} t_i b_i b_i^T\right] v \\
&= \tfrac{1}{2} v^T A(t) v \\
&= \tfrac{1}{2} f^T A^{-1}(t) f.
\end{aligned}$$

Ajoutons aussi qu'il est d'usage de mettre des contraintes de limites pour les valeurs des t_i. Ces t_i devront être compris entre $\underline{t} > 0$ (très petit) et $\bar{t}$. Puisque le problème CTT se traduit comme le problème de la minisation de l'énérgie potentielle, nous obtenons finallement

$$\begin{cases} \min\limits_{t,s} & s \\ \text{s.c.} & f^T A^{-1}(t) f \leq s \\ & \underline{t} \leq t_i \leq \bar{t}, \quad i = 1, \ldots, n. \end{cases}$$

Nous avons déjà rencontré l'inégalité $f^T A^{-1}(t) f \leq s$ dans la section 2.1.4 et nous avons montré qu'elle est équivalente à un ensemble de contraintes linéaires et du cône du second ordre.

B.0.3 Optimisation d'un portefeuille

Nous allons considérer un problème classique de portefeuille concernant n stocks de marchandises que l'on possède sur une période. Pour $i = 1, \ldots, n$, nous désignons par x_i la quantité de la marchandise i disponible au début (et

pendant) la période, et par p_i le prix d'une unité de la marchandise i, de sorte que le revenu r s'exprime par $r = p^T x$. La variable d'optimisation est le vecteur de portefeuille $x \in \mathbb{R}^n$. Les hypothèses de bases sont $x_i \geq 0$, $i = 1, \ldots, n$ et $x_1 + \ldots + x_n = 1$ (budget total unitaire).
Etant donné que les prix des marchandises sont soumis à des variations au cours de la période, il est utile de considérer que la variable p est un vecteur aléatoire de $\mathbb{R}^n$; plus précisément, nous utiliserons pour p un modèle gaussien de moyenne $\bar{p}$ et de matrice de covariance Σ que nous supposons connues. Par conséquent, le revenu associé au portefeuille $x \in \mathbb{R}^n$ est la variable aléatoire normale $r = p^T x$ de moyenne $\bar{r} = \bar{p}^T x$ et de variance $\sigma_r = x^T \Sigma x = \|\Sigma^{1/2} x\|^2$. En utilisant des inégalités du cône du second ordre, nous sommes capables d'exprimer des contraintes qui limitent le risque de perte à plusieurs niveaux. Considérons une contrainte de risque de perte d'un niveau non-désiré maximal α

$$\text{Prob}(r \leq \alpha) \leq \beta, \tag{B.10}$$

où β est un réel non-négatif fixé qui est, typiquement, inférieur à $1/2$. Tout comme nous l'avons fait dans le problème linéaire robuste statistique de la section 2.2.1, nous pouvons réexprimer cette contrainte en utilisant la fonction de répartition Φ d'une variable aléatoire normale de moyenne nulle et de variance unitaire. A partir de (B.10) nous avons

$$\begin{aligned} \text{Prob}\left(\tfrac{r-\bar{r}}{\sqrt{\sigma_r}} \leq \tfrac{\alpha-\bar{r}}{\sqrt{\sigma_r}}\right) \leq \beta \quad &\Leftrightarrow \quad \Phi\left(\tfrac{\alpha-\bar{r}}{\sqrt{\sigma_r}}\right) \leq \beta \\ &\Leftrightarrow \quad \tfrac{\alpha-\bar{r}}{\sqrt{\sigma_r}} \leq \Phi^{-1}(\beta), \\ &\Leftrightarrow \quad \bar{p}^T x + \Phi^{-1}(\beta)\|\Sigma^{1/2} x\| \geq \alpha, \end{aligned}$$

où $\Phi^{-1}(\beta) \leq 0$ si $\beta \leq 1/2$. Il est évident que puisque $\Phi^{-1}(\beta) \leq 0$, cette inégalité constitue une contrainte du cône du second ordre.
Le problème de maximiser le revenu attendu sous la contrainte du risque de perte, avec $\beta \leq 1/2$, peut alors être formulé par le problème SOCP suivant comportant une seule contrainte du cône du second ordre

$$\left\{ \begin{array}{ll} \max & \bar{p}^T x \\ \text{s.c.} & \bar{p}^T x + \Phi^{-1}(\beta)\|\Sigma^{1/2} x\| \geq \alpha \\ & x \geq 0, \quad \sum_{i=1}^n x_i = 1. \end{array} \right.$$

Il existe plusieurs extensions de ce simple problème. Par exemple, nous pouvons imposer plusieurs contraintes de perte, i.e.,

$$\text{Prob}(r \leq \alpha_i) \leq \beta_i, \quad i = 1, \ldots, k,$$

où $\beta_i \leq 1/2$, $i = 1, \ldots, k$, qui exprime le risque (β_i) que l'on serait prêt à courir pour certains niveaux de perte (α_i).
Comme autre variation, nous pouvons tenir compte de l'incertitude dans le modèle statistique $(\bar{p}, \Sigma)$ des prix des marchandises au cours de la période. Supposons que nous ayons N scénarios différents possibles tels que chacun d'entre eux,

disons le k-ème scénario, est représenté par un modèle à distribution gaussienne pour le vecteur aléatoire des prix p_k, dont la moyenne est $\bar{p}_k$ et la matrice de covariance est Σ_k. Une idée est de considérer une approche du pire des cas qui consiste à maximiser le minimum des revenus pour les N scénarios différents, en tenant compte de la contrainte sur le risque de perte pour chaque scénario. En d'autres termes, nous aboutissons au problème SOCP suivant :

$$\begin{cases} \max\limits_{x} & \min\limits_{k} \bar{p}_k^T x \\ \text{s.c.} & \bar{p}_k^T x + \Phi^{-1}(\beta)\|\Sigma_k^{1/2} x\| \geq \alpha, \ k = 1, \ldots, N \\ & x \geq 0, \quad \sum\limits_{i=1}^{n} x_i = 1. \end{cases}$$

Une autre extension autorise d'avoir des cas de pénurie pour les marchandises, i.e. ; $x_i < 0$. Pour pouvoir modéliser de tels cas de figure, nous introduisons les variables x_{long} et x_{short}, avec

$$x_{long} \geq 0, \quad x_{short} \geq 0, \quad x = x_{long} - x_{short}, \quad \sum_{i=1}^{n} x_{short,i} \leq \eta \sum_{i=1}^{n} x_{long,i},$$

où la dernière contrainte impose à la quantité totale de marchandises en cas pénurie de ne pas dépasser une certaine fraction ($0 < \eta < 1$) de la quantité totale de marchandises en période très favorable.

B.0.4 Equilibre d'un système de ressorts linéaires par morceaux.

Considérons un système mécanique constitué de N noeuds situés aux positions $x_1, \ldots, x_N \in \mathbb{R}^2$, où un noeud quelconque i est relié au noeud $i+1$, pour $i = 1, \ldots, N-1$, par un ressort non-linéaire. Les noeuds x_1 et x_N sont fixés à des positions respectives a et b. La tension T_i dans le ressort i est une fonction non-linéaire de la distance entre ses extrémités, i.e., $\|x_i - x_{i+1}\|$:

$$T_i = k(\|x_i - x_{i+1}\| - l_0)_+$$

où $z_+ = \max\{z, 0\}$. Le nombre $k > 0$ désigne la constante de raideur des ressorts et l_0 est leur longueur au repos (sans tension). Sur un noeud quelconque i est attaché un objet de masse $w_i \geq 0$. Le montage est illustré sur la figure B.7.

Le problème est de déterminer la configuration du système, c'est-à-dire les positions $x_2, \ldots, x_{N-1}$ de sorte que la force résultante sur chaque noeud est nulle. Ceci peut être réalisé en cherchant à atteindre le minimum d'énérgie dans le système, ce qui se traduit par le problème suivant

$$\begin{cases} \min & \sum\limits_{i=1}^{N} w_i e_2^T x_i + \sum\limits_{i=1}^{N-1} \phi(\|x_i - x_{i+1}\|) \\ \text{s.c.} & x_1 = a, \ x_N = b \end{cases}$$

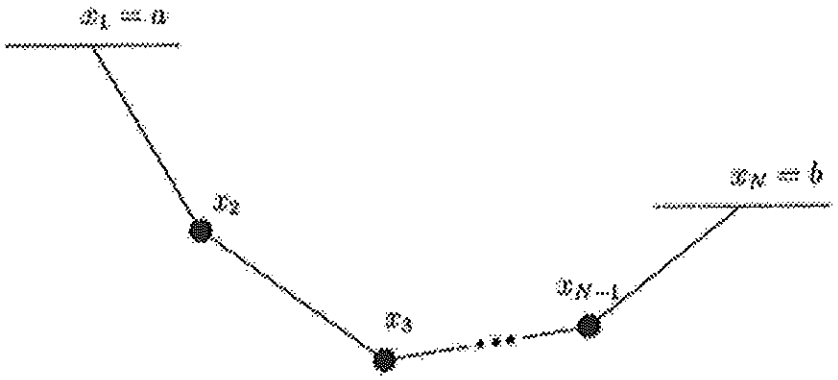


figure B.7: Système de noeuds reliés par des ressorts. Les noeuds x_1 et x_N sont fixés.

où e_2 est le second vecteur de la base canonique de $I\!R^2$ qui pointe vers le haut. Le premier terme de la fonction objectif désigne l'énergie potentielle du système due à la force de gravité, et $\phi(d)$ est l'énergie potentielle emmagasinée par un ressort soumis à l'élongation d. Cette énergie $\phi(d)$ correspond à l'opposé de l'intégrale de la force élastique $F_e(x) = -k(x - l_0)_+$:

$$\phi(d) = \int_0^d k(u - l_0)_+ du.$$

Pour obtenir une expression explicite de cette intégrale considérons les cas où $d \geq l_0$ et $d < l_0$.
Si $d \geq l_0$, nous séparons l'intégrale de 0 à d en deux intégrales de la façon suivante

$$\int_0^d k(u - l_0)_+ du = \int_0^{l_0} \underbrace{k \max\{u - l_0, 0\}}_{=0} du + \int_{l_0}^d k \max\{u - l_0, 0\} du$$

$$= \int_{l_0}^d k(u - l_0) du = \frac{k(d - l_0)^2}{2} = \frac{k(\max\{d - l_0, 0\})^2}{2}.$$

Si $d < l_0$, nous avons

$$0 \leq \phi(d) \leq \int_0^{l_0} \underbrace{k \max\{u - l_0, 0\}}_{=0} du = 0$$

c'est-à-dire, $\phi(d) = 0 = \frac{k(\max\{d - l_0, 0\})^2}{2}$.
Nous avons donc dans les deux cas : $\phi(d) = \frac{k(d - l_0)_+^2}{2}$. Nous pouvons alors reformuler le problème en fonction de cette expression pour $\phi(d)$

$$\begin{cases} \min & \displaystyle\sum_{i=1}^{N} w_i e_2^T x_i + (k/2)\|t\|^2 \\ \text{s.c.} & \|x_i - x_{i+1}\| - l_0 \leq t_i,\ i = 1, \ldots, N-1 \\ & 0 \leq t_i,\ i = 1, \ldots, N-1 \\ & x_1 = a,\ x_N = b, \end{cases}$$

avec $t = (t_1; \ldots; t_{N-1})$. La fonction objectif peut être rendue linéaire en substituant $\|t\|^2$ par y et en ajoutant la contrainte hyperbolique

$$\|t\|^2 \leq y \Leftrightarrow \left\| \begin{bmatrix} 2t \\ 1-y \end{bmatrix} \right\| \leq 1 + y$$

permettant d'obtenir un problème SOCP.

B.1 Tests numériques avec SeDuMi.

Le tableau qui suit résume les résultats obtenus sur des données se rapportant à divers problèmes SOCP en utilisant SeDuMi. Parmi eux, les problèmes nb, nb_L1, nb_L2 et nb_L2_bessel correspondent à des problèmes de synthèse d'un tableau d'antennes, et truss1, truss5, truss8 se rapportent à des problèmes de conception d'une topologie en trellis. Les autres problèmes correspondent à des exemples non décrits ici.[3] Pour chacun des problèmes, le tableau reprend le nombre de contraintes (m), le nombre total de variables (n), le nombre de contraintes du cône du second ordre ($SOCP$), le temps de calcul (t) et le nombre d'itérations pour atteindre l'optimum ($iter$).

nom du problème	m	n	$SOCP$	t	iter
truss1	6	25	7	3	11
truss5	208	3301	34	18	21
truss8	496	1191	34	117	22
nql30	3680	6302	900	78	17
nql60	14560	25202	3600	177	16
nb	123	2383	793	134	18
nb_L1	915	3176	793	162	18
nb_L2	123	4195	839	645	19
nb_L2_bessel	123	2641	839	338	19
qssp30	3691	7566	1891	95	20
qssp60	14581	29526	7381	584	22

[3] Problèmes provenant de la librairie DIMACS :
http ://dimacs.rutgers.edu/Challenges/Seventh/Instances/. Ceux-ci ont été testés avec Matlab 6 sur Pentium II, 300 Mhz.

Bibliographie

[1] F. Alizadeh. Semidefinite and second order cone programming. Programming Seminar, Fall 2001, lecture 12, pages 4-5.

[2] F. Alizadeh. Semidefinite and second order cone programming. Programming Seminar, Fall 2001, lectures 9, 10, 11.

[3] F. Alizadeh and D. Goldfarb. Second order cone programming. Rutgers Center for Operations Research, 2001.

[4] F. Alizadeh and S.H. Schmieta. Extension of primal-dual interior point algorithms to symmetric cones. Rutgers Center for Operations Research, 1999.

[5] A. Ben-Tal and A. Nemirovski. Lectures on Modern Convex Optimization, *Analysis, Algorithms and Engineering Applications*. MPS-SIAM Series on Optimization, 2001.

[6] S. Boyd, L. El Ghaoui, E. Feron, and V. Balakrishnan. Linear Matrix Inequalities in System and Control Theory. SIAM, Philadelphia PA, 1994.

[7] M.R. Hestenes. Optimization Theory, *The finite dimensional case*. John Wiley and Sons, New-York, 1975.

[8] J.Farault and A.Koranyi. Analysis on symmetric cones. Oxford University Press, 1994.

[9] Ulf T. Jönsson. A lecture on the s-procedure. Division of Optimization and Systems Theory Royal Institute of Technology 10044 Stockholm, Sweden, 2001.

[10] Miguel Sousa Lobo. Applications of second order cone programming. citeseer.nj.nec.com/511171.html, 1998.

[11] J-J. Strodiot. Interior-Point Methods in Convex Optimization. Département de mathématiques, Université de Namur, Belgique, 2002.

[12] J. Sturm. Implementation of interior point methods for mixed semidefinite and second order cone optimization problems. Department of Econometrics, Tilburg University, The Netherlands, August 2002. Research Report.

[13] J.H. Wilkinson. The algebraic eigenvalue problem. Oxford University Press, 1965.

[14] V.A. Yakubovich. Minimization of quadratic functionals under quadratic constraints and the necessity of a frequency condition in the quadratic criterion for absolute stability of nonlinear control systems. In Soviet Math. Dokl., volume 14(2), pages 593–597. 1973.